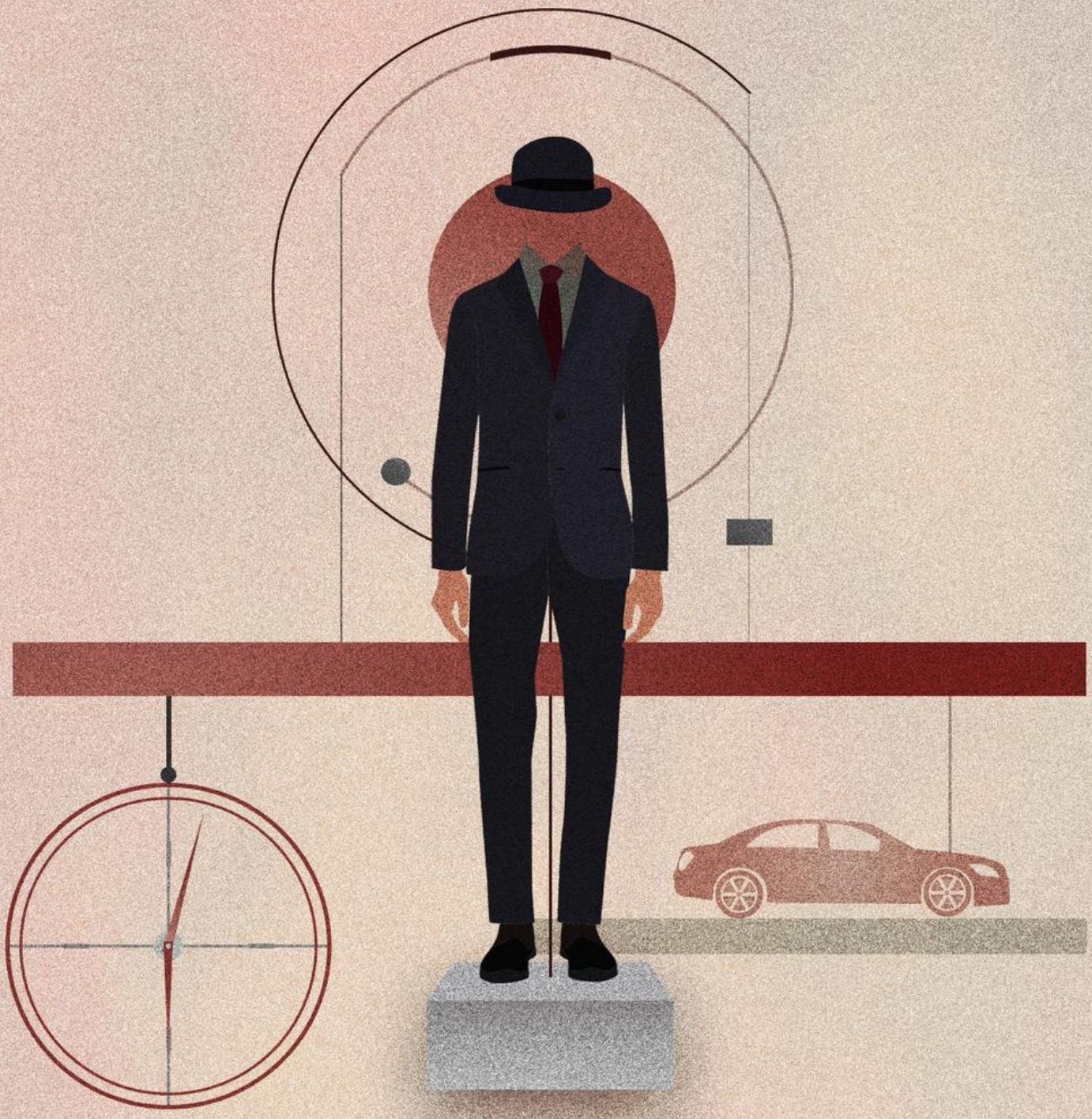
Dynamic traffic equilibria with route and departure time choice

Frascaria, Dario

2021

document version
Publisher's PDF, also known as Version of record



Dynamic traffic equilibria with route and departure time choice

Dario Frascaria

Dynamic traffic equilibria with route and departure time choice

Dario Frascaria

The author was supported by the Netherlands Organisation for Scientific Research (NWO);
grant details: number 614.001.510
title *"Understanding dynamic aspects of traffic"*

The illustration on the cover was made by Chorong Yoo.

ISBN: 978-90-361-0665-8

VRIJE UNIVERSITEIT

Dynamic traffic equilibria with route and departure time choice

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. C.M. van Praag,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de School of Business and Economics
op dinsdag 26 oktober 2021 om 9.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door

Dario Frascaria

geboren te Foggia, Italië

promotor:	prof.dr.	L. Stougie
copromotor:	dr.	N.K. Olver
promotiecommissie:	prof.dr.	T. Harks
	dr.ir.	T.S. Oosterwijk
	prof.dr.	B. Peis
	prof.dr.	J. Rouwendal
	prof.dr.	G. Schäfer
	dr.	V.A.C. Van den Berg

Contents

Chapter 1

Introduction

1.1 Topic of the thesis

Traffic congestion is a phenomenon widespread in the world with negative impacts on several aspects of the society, environment and economy. Several measures have been taken by authorities to alleviate this phenomenon, such as, to name a few, improvements of road infrastructure (junction improvements, separate lanes for specific vehicle groups etc.), policies to reduce the demand of roads (road pricing, parking costs, number plate restrictions etc.) or policies to increase the supply of roads (increase in the number of lanes or creation of new routes). However, sometimes these measures can be ineffective or even counter-productive, as in the case of creating a new route, which can worsen the congestion instead of improving it [Braess, 1968, Kolata, 1990].

One of the key causes of traffic congestion relies on people behavior. People do not coordinate their actions in order to avoid the creation of traffic jams but rather behaves *selfishly*, in the sense that everyone makes decisions that favor himself and not the community. The goal of every road user is, indeed, to arrive at his destination in the most favorable way, e.g. in the fastest or cheapest way. This means that each user makes his own choices, for example regarding the route or the departure time, with no regard for the consequences of these on others. However, the congestion and delays encountered by a user depend partly on the behavior of other individuals and thus users have to adapt their choices according to others' choices. Since different user choices affect each other, the following question can be posed:

Can a user make choices that turn out to be the best possible ones?

This question finds an answer in the notion of *Nash equilibria*, or just *equilibria* [Nash, 1951]. These are situations where nobody has an incentive to unilaterally change their current choices; in other words, situations where everyone is satisfied with his own choices and would not benefit from changing them.

Traffic equilibria can be *static* or *dynamic*. In static equilibria, also called *Wardrop user equilibria* due to the work of Wardrop [1952], the transit time on a road depends on the total amount of users traversing it and does not vary over time, thus establishing interactions among all the users that take the same road. Static equilibria thus do not possess the dimension of time. Conversely, in dynamic equilibria the congestion and transit

time of a road changes over time, depending on the amount of traffic that has crossed it up to that specific moment and independently on the pattern of traffic that will cross it at a later time.

Clearly traffic conditions vary over time and, hence, dynamic equilibria are more accurate than the static ones in representing traffic behavior. For this reason in this thesis we focus exclusively on dynamic equilibria.

From an academic point of view the topic of traffic congestion has received substantial attention from many different communities, each with their own perspective. However, in spite of all the effort that has been put in studying the problem, our theoretical understanding of it is still very limited. For instance, we are not yet able to predict with any confidence where and when congestion happens and, as a consequence, we do not know how to prevent it.

Several models have been introduced for forecasting traffic congestion (see [Peeta and Ziliaskopoulos, 2001] for a survey). Some of them are empirical, where traffic flow is forecast by running and observing simulations on these models. These are very flexible in modeling the behavior of individual vehicles but cannot be thoroughly investigated. Other models are not completely realistic but are used to identify, analytically, the properties that characterize the equilibrium behavior. When choosing a model there is always a trade-off between accuracy and efficiency.

In this thesis we focus on dynamic mathematical models where all the vehicles are identical and where the delays on a road are not caused by traffic lights, car accidents etc., but depend exclusively by the routing choices of the users. These models are built on models used in two different communities: the optimization community and the transportation economics community. They are not completely realistic but they capture the essence of the dynamics behind traffic congestion. In fact, the intent of this thesis is to identify and illustrate concepts regarding traffic behavior and not to represent actual traffic conditions. This aspect is especially important for transportation economists who are interested in investigating the effects of transportation policies. Since traffic involves many types of vehicles, locations, demands and supplies, dealing with fully accurate models is analytically intractable.

To be more specific, we represent traffic dynamics utilizing the *Vickrey bottleneck congestion model* [Vickrey, 1969] (also known as the fluid queuing model). Here traffic flow is treated as a fluid in a pipeline system and each vehicle is represented by an infinitesimally small particle. This implies that the model can be used also when the number of vehicles is gigantic. To each link of the network are associated two parameters, one that indicates how much mass of vehicles can travel in parallel along the link and the other indicates the length of link. If too many vehicles try to enter the link at the same time, some of them pile up at the entrance thus forming a queue (see Figure 1.1). This queue has no physical dimension, in the sense that it can grow indefinitely and that the vehicles joining the queue arrive at the back of it instantaneously and independently of the size of it. A queue on a link does not interfere with traffic on other links and, hence, the total travel time required to traverse a link depends only on its length and on the waiting time spent in its queue.

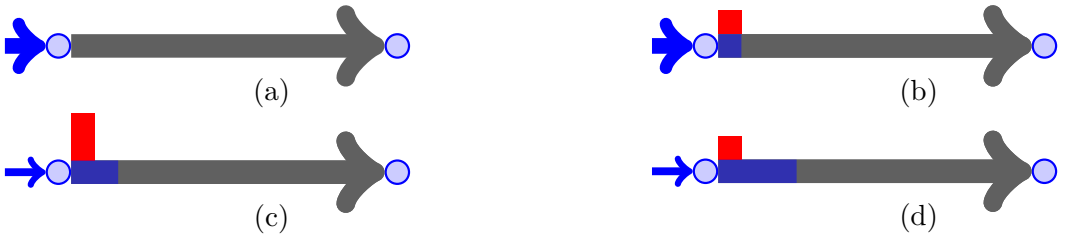


Figure 1.1: Chronological sequence showing the dynamics of the Vickrey bottleneck congestion model. In blue the moving flow and in red the flow waiting in the queue.

As mentioned earlier, this thesis connects the optimization and transportation economics literatures and tackles questions that are of interest for both communities, such as the following:

- Does an equilibrium always exist?
- Can there be different equilibria?
- How can an equilibrium behavior be computed?
- How can one set tolls on roads so that, in an equilibrium, there is no congestion and social welfare is maximized?

Outline of the chapter. We first discuss the optimization literature on dynamic equilibria, in Section 1.2. Then, in Section 1.3, we discuss the perspective of the transportation economics literature related to it. Finally, in Section 1.4 we give an outline of the thesis.

1.2 The perspective from optimization

In the past decades dynamic equilibria have been receiving increasing attention from the optimization community. This line of literature focuses on settings with homogeneous users, given inflow rates and general networks and contains results regarding existence, uniqueness, structural properties, computational difficulties and inefficiency of the equilibria.

One of the earliest works on dynamic equilibria in general networks is the one of Janson [1991], who modeled the problem as a mathematical program. Since the work of Janson, mathematical programming formulations received considerable examination (see Peeta and Ziliaskopoulos [2001] for a review). In these formulations, both the number of users and the time period are discretized and, furthermore, some limitations are present, as the inadequacy to model delay functions that well represent traffic dynamics.

Friesz et al. [1993] were the first to formulate and analyze dynamic equilibria using *variational inequalities*[1]. Since then variational inequalities formulations have received

[1]Inequalities involving functionals, which have to be solved for all possible values of a given variable.

substantial attention (see [Friesz and Han, 2019] for a survey). They are mostly path-based and, consequently, often computationally intractable. Their results thus concern principally the existence and uniqueness of the equilibrium.

Recently, dynamic equilibria have been analyzed from the perspective of algorithmic game theory. Most of the results concern *nonatomic* settings, i.e. setting where there are an infinite number of users, each controlling an infinitesimally small amount of flow. The community has focused more on nonatomic settings mainly because their mathematical structure seems to be more amenable to giving insights and because it's an excellent approximation. Example of atomic settings, however, are the following [Hoefer et al., 2011, Cao et al., 2017, Ismaili, 2017, Scarsini et al., 2018, Werth et al., 2014]

Anshelevich and Ukkusuri [2009] considered a discrete time setting with nonatomic players and general delay functions. They showed that, for flow-dependent delays function obeying FIFO and networks with a single origin-destination pair, an equilibrium always exists and can be computed efficiently while, for networks with multi origin-destination pairs, equilibria may not exist.

All of the works discussed so far do not provide any insights into the structure of the equilibria. The first model to shed light on these is the one of Koch and Skutella [2011], based on the Vickrey bottleneck congestion model [Vickrey, 1969] and the flow over time models [Ford and Fulkerson, 1958]. The authors considered a setting with a single origin-destination pair where users are released over time at a constant rate, called *inflow rate into the network* (or simply *inflow rate*). They provided an elegant characterization of the equilibrium based on the derivatives (with respect to time) of the distance between the origin and any other vertex. Furthermore, they showed that the equilibrium can be decomposed into static flows, with specific structures, that they called *thin flows with resetting*. Using these static flows they developed an algorithm to construct an equilibrium. They did not provide, however, any bounds on the time complexity of the algorithm, i.e. on the amount of time required to run the algorithm. The work of Koch and Skutella [2011] serves as a basis for all the work listed in the remainder of this review.

Cominetti et al. [2015] proved the existence of thin flows with resetting and they turned the algorithm of Koch and Skutella [2011] in a proof of existence of equilibria for the case of a piecewise constant inflow rate. They also showed that, in the case of piecewise constant inflow rate, the rate of change of the queues in an equilibrium are, under some technical conditions, unique. Moreover the authors, using variational inequalities, extended the existence result also to the case of inflow rate functions belonging in the L^p space and also to multiple origin-destination settings.

Kaiser [2020] showed that computing a thin flow with resetting is a linear complementarity problem and that can be solved in polynomial time in series-parallel networks. The complexity on general networks remains unknown.

Cominetti et al. [2017] showed that the equilibrium flow reaches a *steady state* in finite time when the inflow rate into the network is a constant not bigger than the min-cut capacity of the network. A steady state is an unbounded time interval in which the queue delays and the transit times do not change anymore, and thus a state where the queue lengths are frozen. They also prove that the equilibrium can have an exponential number of phases, making the complexity of the Koch-Skutella algorithm to construct an equilibrium at least exponential.

Some works focused on the inefficiency of the equilibria for different objectives [Koch and Skutella, 2011, Bhaskar et al., 2011, Correa et al., 2019]. The inefficiency of the equilibria is measured with the *price of anarchy* [Koutsoupias and Papadimitriou, 1999, Papadimitriou, 2001], that is defined as the worst possible ratio between the total cost of an equilibrium and the one of a *social optimum* – an optimal behavior that minimizes the average cost of all the agents (i.e. maximizes the social welfare) but that, usually, is not an equilibrium.

Macko et al. [2010] studied a new type of Braess's paradox appearing in this model. Namely they consider the maximum experienced latency of a flow particle and they conjecture a necessary and sufficient condition for the paradox.

Recently the model has been extended to include further traffic elements. Sering and Skutella [2018] generalized the thin flows with resetting to multiple origin-destination pairs networks, where the outflow of every origin is routed to the destinations proportionally with respect to a global pattern. Their model, however, assumes that flow particles are essentially equal, apart from their origin or destination. In fact, once they meet at any location, flow particles with different origins and or destinations cannot be distinguished. Sering and Vargas Koch [2019] extended the model by dropping the assumption that queues can grow indefinitely. Instead, they introduced an upper bound on total amount of flow present on a link at any time. When the flow exceeds this value, the congestion propagates backwards to the preceding links. This dynamic is called *spillback*. The authors generalized the thin flows with resetting to this setting, showed existence of the equilibria and suggested an algorithm for their computation.

All the aforementioned studies assumed the flow particles to have complete information on the state of the network for all points in time. This means that all the road users know all others' route choices and can exactly foresee the travel times of all the paths. This is justified by assuming that the equilibrium behavior repeats itself as a daily routine and hence that users have learnt each others' choices over several trips. Graf and Harks [2019] introduced a setting where this does not occur and where, instead, users have knowledge only on the current network conditions, i.e. on the current queuing delays. Therefore users do not predict how the queue lengths will change during their trip but, instead, they choose, at any intersection, the shortest route according to the current queues. The authors proved existence and nonuniqueness [Graf and Harks, 2019, Graf et al., 2020], provided an upper bound on the price of anarchy [Graf and Harks, 2020b] and presented complexity results on the computation of the equilibrium [Graf and Harks, 2020a].

In spite of the elegance of the Koch-Skutella model and of all the attention that it has received, many fundamental problems remain unsolved and several apparently obvious properties remain hard to demonstrate. An example is given by the *monotonicity conjecture* found in [Correa et al., 2019], which relates the travel time of particles to the inflow rate. Other similar conjectures can be found in Chapter 4 of this thesis.

1.3 The perspective from transportation economics

In this section we discuss the perspective of the transportation economics literature connected to the work discussed in Section 1.2.

Transportation economists focus especially on settings where the users' interactions are more intricate. Their motivations rely on the basis that traffic congestion is not just a physical phenomenon but it occurs as a consequence of people's decisions and these vary from individual to individual, according to personal factors or preferences. Their models, therefore, incorporate other aspects of user utility besides route transit times and, often, exhibit heterogeneous groups of users. This goes to the detriment of the network topology which is often very simple, like a single-link or multiple parallel-links network.

A very standard setting, motivated by morning rush-hour traffic, is the *scheduling cost* one. This was introduced in the *bottleneck model* together with the bottleneck congestion dynamic [Vickrey, 1969]. Since then it has been elaborated in many works (see [Small, 2015, Li et al., 2020]). Here users desire to arrive at their destination at a particular time and their total disutility, called **travel cost**, consists of two components:

- a **journey time cost**, precisely the time between the departure of the user from the origin and their arrival at the destination, scaled by a factor α (the value of time); and
- a **scheduling cost**, based on the arrival time of the user compared to a fixed desired arrival time. We write $\rho(\theta)$ for the scheduling cost associated with arriving at time θ.

A very standard choice for a scheduling cost function is

$$\rho(\theta) = \begin{cases} \beta(T^* - \theta) & \text{if } T^* \geq \theta \\ \gamma(\theta - T^*) & \text{otherwise} \end{cases} \tag{1.1}$$

where T^* represents the desired arrival time and where $0 < \beta < \alpha < \gamma$ (it is very bad to be late, but time spent in the office early is better than time spent in traffic). This scheduling cost function is represented in Figure 1.2.

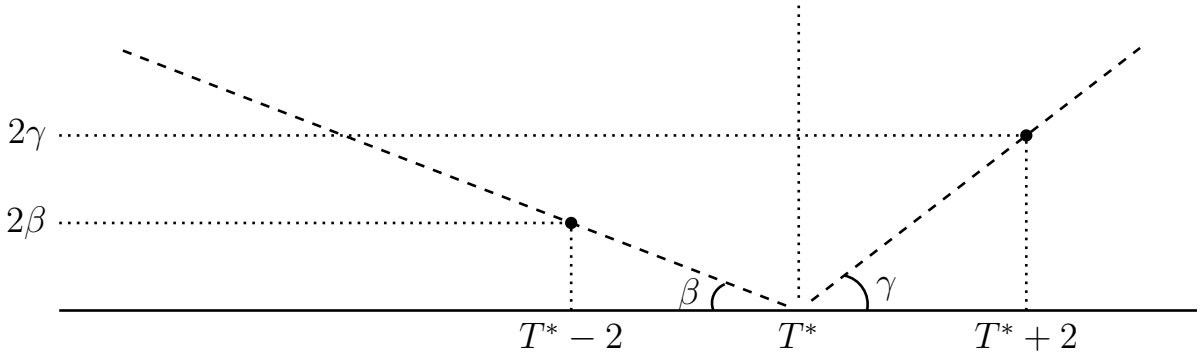
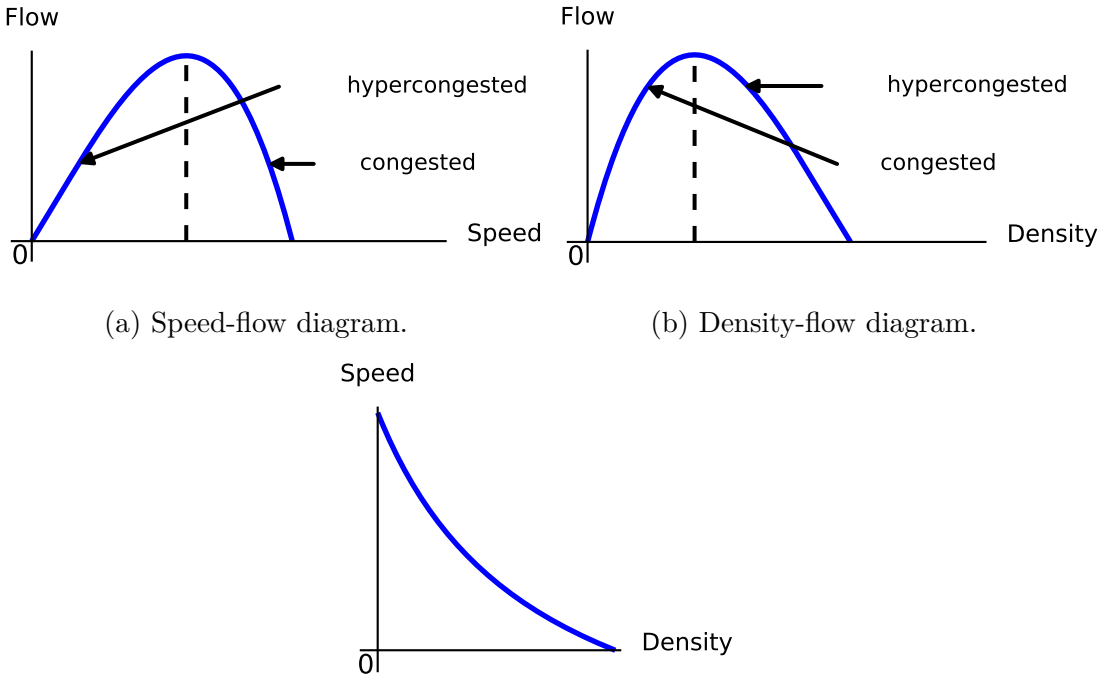


Figure 1.2: The horizontal axis represents the arrival time. The dashed line indicates the scheduling cost function.

With this setting, in an equilibrium users choose not only their routes but also their departure times. This means that inflow rate into the network is not given but depends on the aggregate choices of all the users. We will use this setting in most of the thesis, namely in Chapters 3 to 6.

While the optimization literature investigates more the computational difficulties and structural properties of the equilibria, this line of research investigates more policies and their effects. An example of such policies is *pricing*. This is used as a tool to persuade people to make socially efficient choices and thus increasing the social welfare. Since the first works of Pigou [1920], Knight [1924], Walters [1961], Vickrey [1963], this subject has received substantial attention from the transportation economics community (see [Lindsney and Verhoef, 2001] for a survey). However, for the *optimal tolls*, i.e. the ones that induce an equilibrium that maximizes the social welfare, this line of literature does not provide an elegant characterization but rather contains iterative toll adjustment procedures (trial-and-error implementation schemes) and dynamic equilibrium simulators that compute them (as, for example, [De Palma et al., 2005] or [Yang et al., 2004, Meng et al., 2005]). A novel characterization for these tolls is given in Chapter 3 of this thesis.

Another difference is given by the fact that, although both communities use the Vickrey bottleneck congestion model, transportation economists consider also different kinds of congestion delay functions (see [Lindsey and Verhoef, 1999] for a survey). The interest in different delay functions resides in representing empirically observed behavior, like the one of *hypercongestion*. A classic way to model traffic congestion, of particular importance in modelling highway traffic, is by utilizing three variables that represent *traffic speed*, *traffic density* and *traffic flow*, measured, respectively, in distance per unit of time, number of vehicles per unit of length and number of vehicles per unit of time [Walters, 1961]. Traffic flow equals the product of speed and density and these, usually, are in a negative linear relationship: speed decreases as density increases (Figure 1.3c). Traffic flow is thus an inverse u-shaped function of speed and density (Figures 1.3a and 1.3b) and the same value, when not maximized, can be obtained by two different combinations of speed and density. The flow is said to be *congested* when it stems from high speed and low density and *hypercongested* when it stems from low speed and high density. We will discuss hypercongestion with Vickrey bottleneck congestion dynamics in Chapter 5 of this thesis.



(a) Speed-flow diagram.

(b) Density-flow diagram.

(c) Speed-density (or fundamental) diagram.

Figure 1.3: Fundamental diagrams of traffic flow.

1.4 Contributions and outline of the thesis

This thesis combines models of the optimization and the transportation economics literatures.

The methodology we use follows the treatments in the optimization literature. More precisely, we study equilibria by extending the model introduced by Koch and Skutella [2011] and further elaborated by Cominetti et al. [2015] and we analyze traffic streams utilizing *flows over time*. These are discussed in Chapter 2.

We mainly focus on homogeneous settings with arbitrary network topologies using Vickrey bottleneck congestion and scheduling costs. Our intent is thus to extend the classical Vickrey bottleneck model to general networks.

In Chapter 3 we present a combinatorial algorithm for computing the social optimum – the behavior that minimizes the average total cost of all users (i.e., maximizing the social welfare) – in a setting with scheduling costs. Here we consider any scheduling cost function. Moreover, we show how to set optimal tolls, i.e. tolls that induce an equilibrium that corresponds to a social optimum. We will use these tolls in Chapter 5. Furthermore, we also show that our construction of optimal tolls is applicable to the *exogenous demand model*, i.e. the model where users are released at constant rate into the network. This model will be considered in Chapter 7.

In Chapter 4 we define the model of the *dynamic equilibria with endogenous departure*

time choice, equilibria where users choose both their route and their departure time. For the latter we consider a setting with scheduling costs, similar to the one of Chapter 3 but restricted to convex functions. This model is applied in Chapters 5 and 6.

Additionally, we show existence and uniqueness under two very plausible conjectures about continuity and monotonicity of equilibria. Finally, we demonstrate how this model generalizes the one where users do not choose the departure time and that will be treated in Chapter 7.

In Chapter 5 we study hypercongestion. Hypercongestion is the phenomenon when an increase of traffic densities corresponds to a decrease of traffic throughput. It is well understood at the single-link level but it has also been observed in a macroscopic form at the level of traffic networks; for instance, in morning rush-hour traffic into a downtown core. For this reasons, transportation economists employ link delay functions that exhibit hypercongestion. In this chapter we show that macroscopic hypercongestion can occur as a purely emergent effect of dynamic equilibrium behavior on a network, even when it is not present in the link delay function.

In Chapter 6 we study *the traffic corridor problem*, initialized by Arnott [2001], Arnott and DePalma [2011a]. Here the network is a corridor connecting a continuum of residential locations to a single point, the *central business district (CBD)*, and that is subject to flow congestion. Arnott and DePalma [2011a] studied the equilibrium behavior using the kinematic model [Lighthill and Whitham, 1955, Richards, 1956] and they determine important properties of an equilibrium, if it exists. However they argue that an equilibrium does not always exist for their model. Akamatsu et al. [2015] studied a version of the problem where the corridor is not continuous but is modeled as a series-link network[2] and where the Vickrey bottleneck congestion model is used. They showed that model can be formulated as a linear complementarity problem and, by utilizing Kakutani's fixed point theorem, they proved existence. Furthermore, for a range of parameter choices that excludes the homogeneous case, they proved uniqueness. It is worth noting that they discretized time to simplify the analysis.

In our work we improve the results of [Akamatsu et al., 2015] by proving existence and uniqueness of the equilibrium (without discretizing time) for the homogeneous case, we provide a description of the structure of the equilibrium and we develop a polynomial time algorithm for computing the equilibrium flow.

In the final chapter, Chapter 7, we consider *dynamic equilibria with exogenous departure time choice* – equilibria where users do not choose their departure time but are released at a constant inflow rate. Here the cost of a user is just its journey time. This is the model that has received most attention from the optimization community (see Section 1.2).

Cominetti et al. [2017] studied the equilibrium behavior under an assumption on the inflow rate into the network and showed that, in finite time, the equilibrium reaches a steady state, a state where the behavior of the equilibrium stabilizes, in the sense that queue delays and transit times do not change anymore. Their proof is based on a potential function that is monotone along the evolution of the equilibrium.

In this chapter we provide a more general definition and characterization of steady

[2] A graph that consists of a series of arcs.

states and we drop the assumption on the inflow rate into the network. This characterization leads to the definition of a potential function that is quite natural and that generalizes the one used in [Cominetti et al., 2017]. However, we found a network and queue lengths configuration where the potential function is not monotone. We do not know if such configuration is ever achieved by the equilibrium and we conjecture that further insights into the relation between consecutive thin flows with resetting are needed in order to solve this problem.

Figure 1.4 shows the connections among chapters.

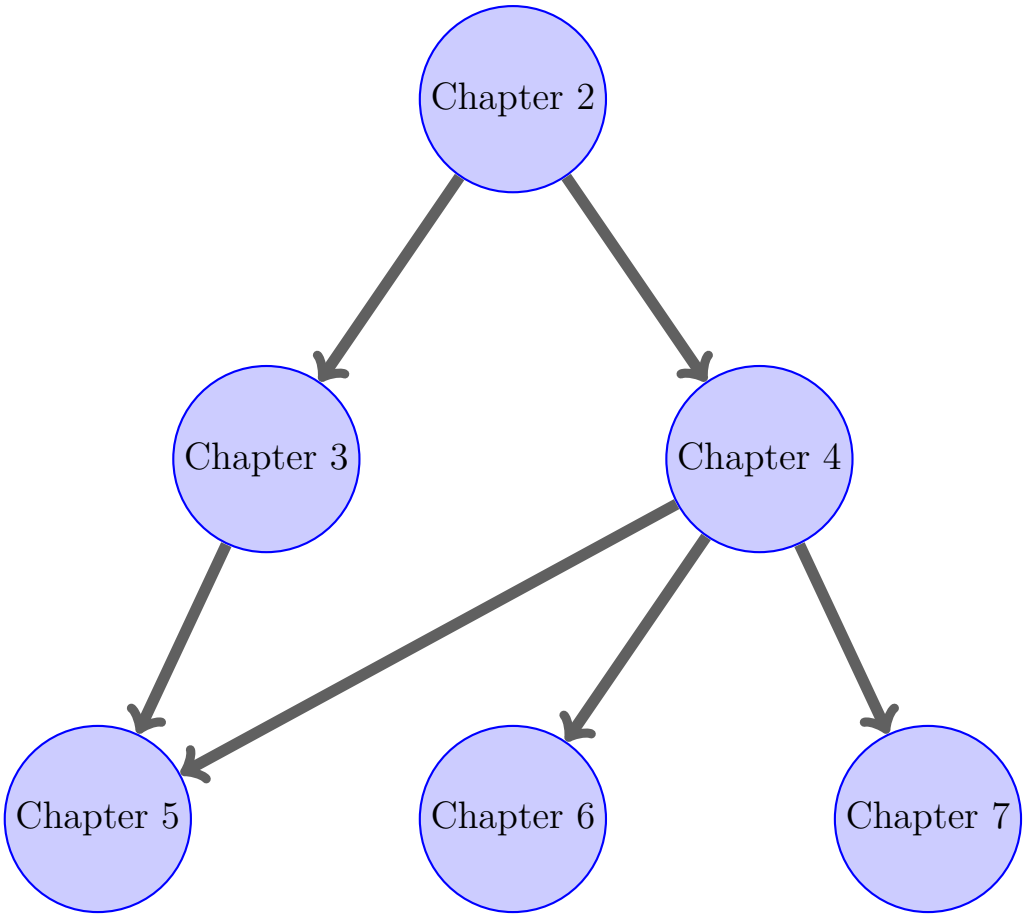


Figure 1.4: Graph showing the connections among the chapters. The solid arc from a chapter A to a chapter B indicates that the results of chapter A are used in chapter B.

Chapter 2

Preliminaries: flows over time

2.1 Introduction

In this chapter we introduce the flow over time models and the notation used in the rest of the thesis.

The study of *flows over time* is a classical one in combinatorial optimization; it began already with the work of Ford and Fulkerson [1958] in the 50s. It is a natural extension of static flows, which associates a single numerical value to each arc, representing the total quantity or rate of flow that can traverse the arc. In a flow over time, a second value associated with each arc represents the time it takes for flow particles to traverse it; the flow is then described by a function on each arc, representing the rate of flow entering the arc as a function of time. We consider flows over time in continuous-time settings; here, a flow over time is characterized by flow rates that indicate how much flow passes a certain point (tail or head of the arc) per unit of time.

We use flow over time models in all the remaining chapters of the thesis.

Outline of the chapter. Before introducing the flow over time formally (in Section 2.4), we introduce the network structure (in Section 2.2) and static flows (in Section 2.3).

2.2 Network structure

The network is described by a given directed graph. We usually write $G = (V, E)$ to indicate a graph G with vertex set V and arc set E. We restrict ourselves to the setting of a common source (or origin), which we denote by s, and a common sink (or destination) which we denote by t. Moreover, we assume that every vertex is reachable from s and that there is no directed cycle with zero transit time. A path that starts at v and ends at w is called v-w-path. We use $\delta^{out}(v)$ to denote the set of arcs with tail v, and $\delta^{in}(v)$ the set of arcs with head v. We also call the arcs in $\delta^{in}(v)$ the *in-going arcs* of v and the arcs in $\delta^{out}(v)$ the *out-going arcs* of v.

2.3 Static flows

Let $\mathbb{R}_+$ indicate the set of nonnegative reals and consider some function $x : E \to \mathbb{R}_+$ that assigns flow values to the arcs of the network. Each arc $e \in E$ is associated with a nonnegative *capacity* ν_e, which indicates how much flow can be assigned to it.

For any vertex $v \in V$, we define the *net flow at* v (denoted ∇x_v) to be the difference between the total amount of flow on the in-going arcs of v and the total amount of flow on the out-going arcs of v. Namely, the net flow at v is the quantity

$$\nabla x_v := \sum_{e \in \delta^{in}(v)} x_e - \sum_{e \in \delta^{out}(v)} x_e.$$

We say that x is a *(static)* s-t-flow if

(i) $\nabla x_v = 0$ for all $v \in V \setminus \{s, t\}$

(ii) $x_e \leq \nu_e$ for all $e \in E$

(iii) $-\nabla x_s \geq 0$.

Think of x as an instantaneous moving flow. The first condition, called *flow conservation*, states that, for any vertex apart from s and t, the amount of flow entering the vertex equals the amount of flow leaving it. The second condition states that the amount of flow on an arc cannot exceed its capacity. Finally the last condition specifies the source and, as consequence of flow conservation, the sink. Furthermore, we say that a (static) s-t-flow has *value* Q if the amount of net flow leaving s equals Q, i.e. if $-\nabla x_s = \nabla x_t = Q$.

2.4 Flows over time

Contrary to static flows, flows over time are functions of time. Here flow particles require a certain amount of time to travel through an arc and different flow particles can enter and exit arcs at different times and still affect each other.

Each arc $e \in E$ has a *capacity* ν_e and a *free transit time* τ_e associated with it (both nonnegative). The capacity indicates how much flow can cross a point of the arc in any instant and the free transit time indicates how long it takes to cross the entire arc when there is no waiting (waiting may be caused by congestion).

We can describe the flow on an arc e by functions $f_e^{in}, f_e^{out} : \mathbb{R} \to \mathbb{R}_+$; $f_e^{in}(\theta)$ denotes the *inflow rate* into arc e at time θ, and similarly $f_e^{out}(\theta)$ the *outflow rate*. We require these functions to be measurable and defined almost everywhere (for the flows that we will consider in this thesis, they are in fact piecewise constant).

The total amount of flow that has entered arc e by time θ is

$$F_e^{in}(\theta) = \int_{-\infty}^{\theta} f_e^{in}(\xi) d\xi$$

and the total amount of flow that has left arc e by time θ is

$$F_e^{out}(\theta) = \int_{-\infty}^{\theta} f_e^{out}(\xi) d\xi \ .$$

Traffic is treated as divisible, and vehicles therefore as a continuous flow, and so these functions can take on arbitrary nonnegative values, without any integrality restrictions. Since flow can exit an arc only if it has previously entered it, we have the following inequality:

$$F_e^{out}(\theta + \tau_e) \leq F_e^{in}(\theta) \qquad \forall e \in E, \theta \in \mathbb{R}. \tag{2.1}$$

Similarly to the static case, for any vertex $v \in V$, we define the *net flow into v at time θ* (denoted $\nabla f_v(\theta)$) to be the difference between the total amount of flow leaving the in-going arcs of v and the total amount of flow entering the out-going arcs of v. Namely, the net flow into v at time θ is the quantity

$$\nabla f_v(\theta) := \sum_{e \in \delta^{in}(v)} f_e^{out}(\theta) - \sum_{e \in \delta^{out}(v)} f_e^{in}(\theta) \ .$$

We use two definitions of flow over time, one used in chapter Chapter 3, where particles wait in queues on the vertices, and one used in all the remaining chapters, where particles wait on queues at the entrance of the arcs. These definitions are conceptually the same. A *s-t flow over time with waiting on the arcs* is defined as a family $(f_e^{in}, f_e^{out})_{e \in E}$ of arc inflows, outflows such that

(i) $\nabla f_v(\theta) = 0$ for any $v \in V \setminus \{s, t\}$ and almost every θ

(ii) $-\int_{-\infty}^{\infty} \nabla f_s(\xi) d\xi \geq 0$

(iii) $f_e^{out}(\theta) \leq \nu_e$ for any $e \in E$ and almost every θ.

One can think of a flow over time as a fluid in a pipeline system. The first condition, called *strict flow conservation*, states that, for any vertex apart from s and t, the flow entering a vertex at any time cannot wait on the vertex but has to immediately leave. The second condition specifies the source and, as consequence of strict flow conservation, the sink. The third condition states that the outflow rate of an arc cannot exceed its capacity.

For an *s-t flow over time with waiting on the vertices* we replace the conditions (i) and (iii) with the conditions

(iv) $f_e^{in}(\theta) \leq \nu_e$ for any $e \in E$ and almost every θ

(v) $f_e^{in}(\theta) = f_e^{out}(\theta + \tau_e)$ for any $e \in E$ and almost every θ.

Condition (iv) states that the inflow rate into an arc cannot exceed its capacity and condition (v) states that, once inside an arc, flow particles have to immediately move towards the head of the arc without the possibility to wait on it.

Furthermore, we call a flow over time *without waiting* if it satisfies all the conditions.

Finally, we say that a *s-t*-flow over time has *value* (or *mass*) Q if the amount of flow leaving s equals Q, i.e. if $-\int_{-\infty}^{\infty} \nabla f_s(\xi) d\xi = \int_{-\infty}^{\infty} \nabla f_t(\xi) d\xi = Q$.

In the following, we present an example of a flow over time.

Example 2.1. Consider a network with $V = \{s, a, b, t\}$, $E = \{sa, sb, ab, bt\}$, $\nu_e = 1$ $\forall e \in E$, $\tau_e = 1$ $\forall e \in E \setminus \{ab\}$ and $\tau_{ab} = 0.5$ (see Figure 2.1).

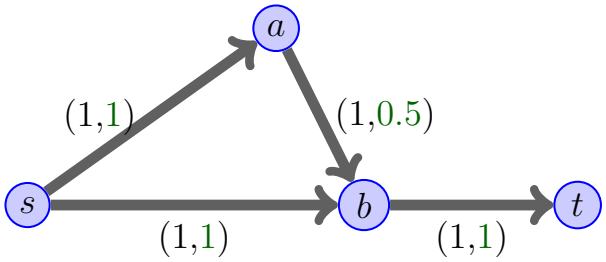


Figure 2.1: Network of Example 2.1. The label of an arc represents the capacity (first element) and the free transit time (second element).

Here the inflow into the network is positive only for one unit of time, from time 1.0 till time 2.0, in which it equals 2. More precisely, the inflow into the network is

$$\nu_0(\theta) := \begin{cases} 2 & \theta \in [1, 2] \\ 0 & \theta \notin [1, 2]. \end{cases}$$

Rather than specifying the inflow and outflow rate for any arc and any point in time, we define this flow over time using the flow particles point of view: Upon entry in the network, each particle of flow split evenly between the paths *sabt* and *sbt* and moves towards the sink as fast as possible, i.e. waiting only when necessary. This means that the particles leave a queue according to the First-In-First-Out (FIFO) principle.

We describe it with a sequence of snapshots of the network (see Figure 2.2):

(a). At time 0.0 the network is empty and there is no flow.

(b). At time 1.0 the first particles of flow appear at the source.

(c). At time 1.5, after half unit of time since departure, the particles that departed first arrive at the middle points of the arcs *sa* and *sb*.

(d). At time 2.0 the inflow into s drops to 0 and the particles that departed first arrive at a and b.

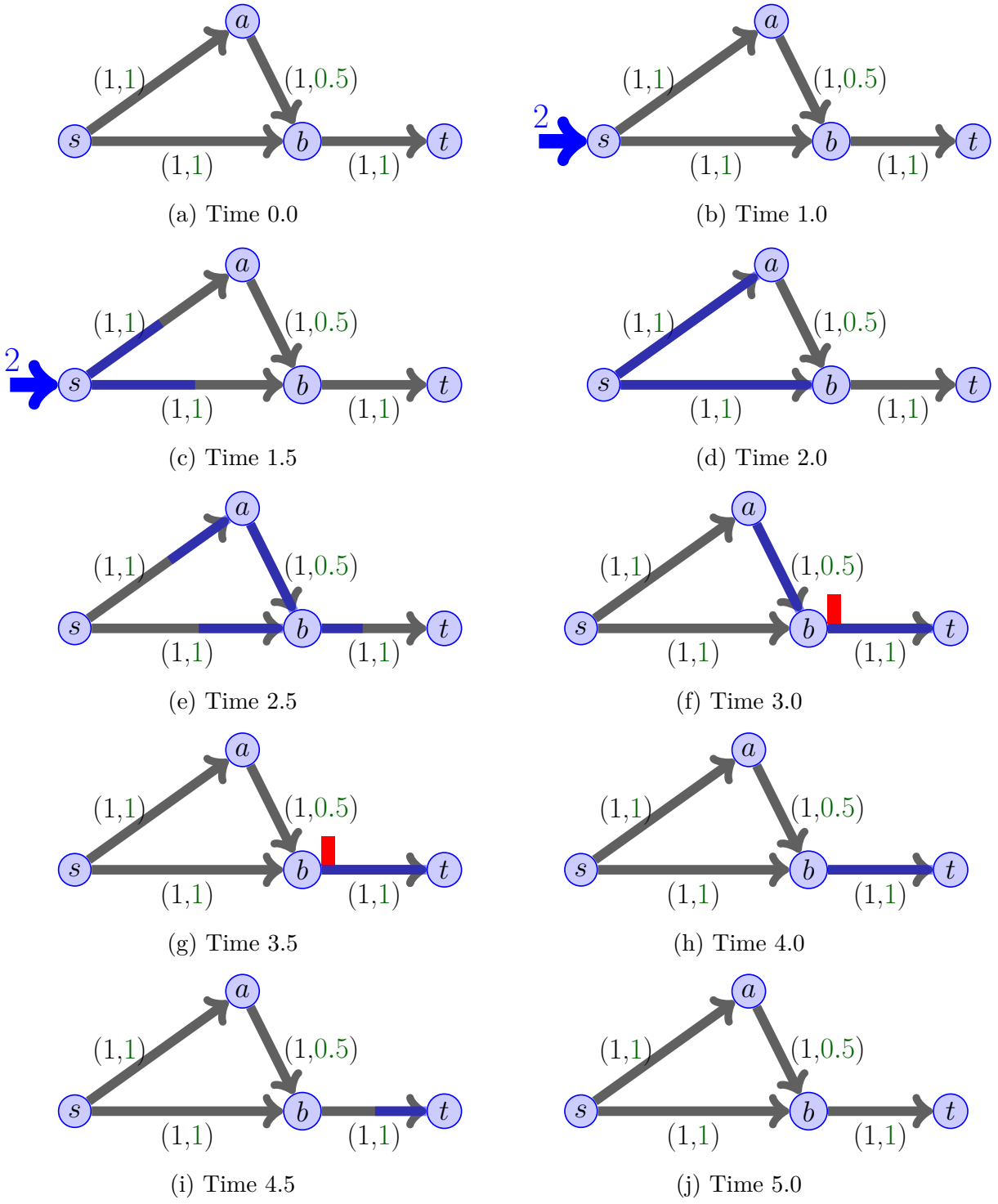
(e). At time 2.5 the inflow into b becomes 2. Since b has only one ongoing arc with capacity 1, some particles have to wait and they will form a queue. The rate of grow of this queue equals 1, the difference between the inflow into bt and its capacity. In Figure 2.2f we represent the situation when they wait on the arc bt rather than on b.

(f). At time 3 the inflow into b becomes 1 and therefore the queue on bt stops growing. The total amount of mass waiting on the arc bt is equal to 0.5.

(g). At time 3.5 the inflow into b becomes 0 and therefore the queue on bt will decrease with rate 1, the difference between the capacity of bt and its inflow.

(h). At time 4.0 the queue on bt is completely depleted.

(i) and (j). In the final two phases, all remaining particles arrive at the sink. The final particle arrives at t at time 5.0.



(a) Time 0.0 (b) Time 1.0

(c) Time 1.5 (d) Time 2.0

(e) Time 2.5 (f) Time 3.0

(g) Time 3.5 (h) Time 4.0

(i) Time 4.5 (j) Time 5.0

Figure 2.2: Chronological sequence of snapshots of the flow over time discussed in Example 2.1. In blue the inflow into the network and the moving flow. In red the nonmoving flow.

Chapter 3

Algorithms for flows over time with scheduling costs

The results in this chapter appear in [Frascaria and Olver, 2020].

3.1 Introduction

Classical optimization problems involving static flows have natural analogs in the flow over time setting (see the surveys [Köhler et al., 2009, Skutella, 2009]). For example (restricting the discussion to single commodity flows), the *maximum flow over time* problem asks to send as much flow as possible, departing from the source starting from time 0 and arriving to the sink by a given time horizon T; this can be solved in polynomial time [Ford and Fulkerson, 1958, 1962, Fleischer and Tardos, 1998]. A *quickest flow* asks, conversely, for the shortest time horizon necessary to send a given amount of flow. Of particular importance for us is the notion of an *earliest arrival flow*: this has the very strong property that simultaneously for all $T' \leq T$, the amount of flow arriving by time T' is as large as possible [Gale, 1959]. Such a flow can also be characterized as minimizing the average arrival time [Jarvis and Ratliff, 1982]. Earliest arrival flows can be "complicated", in that they can require exponential space (in the input size) to describe [Zadeh, 1973], and determining the average arrival time of an earliest arrival flow is NP-hard [Disser and Skutella, 2015]. But they can be constructed in time strongly polynomial in the sum of the input and output size [Baumann and Skutella, 2006].

In this chapter we focus on flow over time in a setting with scheduling costs (see Section 1.3). We allow general scheduling cost functions, though for most of the chapter we focus on strongly unimodal cost functions; these are the most relevant, and this avoids some distracting technical details.

Two very natural questions can be posed at this point. The first is a purely optimization question, with no attention paid to the decentralized nature of traffic.

Question 3.1. How can one compute a flow over time minimizing the average total cost paid by users, i.e., maximizing the social welfare?

From now on, we will call a solution to this problem simply an *optimal flow*.

It is well understood that users typically do not coordinate their actions to induce a flow that minimizes total disutility. There is a huge body of literature (particularly in the setting of static flows [Nisan et al., 2007]) investigating this phenomenon. In the traffic setting, the relevance of an optimal flow represented by an answer to this question comes primarily via the possibility of *pricing*. By putting appropriate tolls on roads, we can influence the behavior of users and the resulting dynamic equilibrium. Thus:

Question 3.2. How can one set tolls (possibly time-varying) on the arcs of a given instance so that an optimal flow is obtained in dynamic equilibrium?

One subtlety is that since dynamic equilibria need not be precisely unique, there is a distinction between tolls that induce an optimal flow as *an* equilibrium, compared to tolls for which *all* dynamic equilibria are optimal. We will call this *weak* and *strong* enforcement of optimality, respectively, and will return to this subtlety shortly. (See Harks [2019] for some related notions of enforcement in a general pricing setting.)

Questions like these are of great interest to transportation economists. However, most work in that community has focused on obtaining a fine-grained understanding of very restricted topologies (such as a single link, or multiple parallel links).

Both of these questions (for general network topologies) were considered by Yang and Meng [1998] in a discrete time setting, by exploiting the notion of *time-expanded graphs*. This is a standard tool in the area of flows over time; discrete versions of all the optimization questions concerning flows over time mentioned earlier can (in a sense) be dealt with in this way. A vertex v in the graph is expanded to a collection (v, i) of vertices, for $i \in \mathbb{Z}$ in a suitable interval, and an arc vw with free transit time τ_{vw} becomes a collection of arcs $((v, i), (w, i + \tau_{vw}))$ (this assumes a scaling so that τ_{vw} is a length in multiples of the chosen discrete timesteps). Scheduling costs are encoded by appropriately setting arc costs from (t, i) to a supersink t' for each i, and the problem can be solved by a minimum cost static flow computation. A primary disadvantage of this approach (and in the use of time-expanded graphs more generally) is that the running time of the algorithm depends polynomially on the number of time steps, which can be very large. Further, it cannot be used to exactly solve the continuous time version (our interest in this chapter); by discretizing time, it can be used to approximate it, but the size of the time-expanded graph is inversely proportional to the step size of the discretization. In the same work [Yang and Meng, 1998], the authors also observe that in the discrete setting, an answer to the second question can be obtained from the time-expanded graph as well. Taking the LP describing the minimum cost flow problem on the time-expanded graph, the optimal dual solution to this LP provides the necessary tolls to enforce (weakly) an optimal flow. (This is no big surprise—dual variables can frequently be interpreted as prices.)

An assumption on the scheduling cost function ρ. Suppose we consider a scheduling cost function ρ in the standard form given in (1.1), but with $\beta > \alpha$. This means that commuting is considered to be less unpleasant than arriving early. A user arriving earlier than time T^* at the sink would be better off "waiting" at the sink before leaving, in order to pay a scheduling cost of 0. Whether waiting in this way is allowed or not depends on

the precise way one specifies the model, but it is most natural (and convenient) to allow this. If we do so, then it is clear that a scheduling cost function ρ can be replaced by

$$\hat{\rho}(\theta) := \min_{\xi \geq \theta} \rho(\xi) + \alpha(\xi - \theta)$$

without changing the optimal flow (except there is no longer any incentive to wait at the sink, and we need not even allow it). Then $\theta \to \hat{\rho}(\theta) + \alpha\theta$ is nondecreasing. From now on, we always assume that ρ satisfies this; we call it the *growth bound* on ρ.

Our results. We give a combinatorial algorithm to compute an optimal flow. Similarly to the case of earliest arrival flows, this flow can be necessarily complicated, and involves a description length that is exponential in the input size.

The algorithm is also similar to that for computing an earliest arrival flow. It is based on the (possibly exponentially sized) path decomposition of a minimum cost flow into *successive shortest paths*. In particular, suppose we choose the scheduling cost function to be

$$\rho(\theta) = \begin{cases} -\alpha\theta & \text{if } \theta \leq 0 \\ \infty & \text{if } \theta > 0. \end{cases} \tag{3.1}$$

Then the disutility a user experiences is precisely described by how much before time 0 they depart; all users must arrive by time 0 to ensure finite cost. This is precisely the reversal (both in time and direction of all arcs) of an earliest arrival flow, from the sink to the source. (By writing the average arrival time objective as the integral over time of the total flow not yet arrived by this time, this exact correspondence is easy to see.) Our algorithm, in this case, is the same (up to the time reversal) as the usual algorithm for earliest arrival flow Fleischer and Tardos [1998].

This also shows that there are instances where all optimal solutions to Question 3.1 require exponential size (as a function of the input encoding length), since this is the case for earliest arrival flows [Zadeh, 1973].

Despite the close relation to earliest arrival flows, the proof of optimality of our algorithm is rather different. A key reason for this is the following. As mentioned, earliest arrival flows have the strong property that the amount of flow arriving before a given deadline T' is the maximum possible, *simultaneously for all choices of* T' (up to some maximum depending on the total amount of flow being sent). This implies that an earliest arrival flow certainly minimizes the average arrival time amongst all possible flows [Jarvis and Ratliff, 1982], but this is a substantially stronger property. A natural analog of this stronger property in our setting would be to ask for a flow for which, simultaneously for any given cost horizon $C' \leq C$, the amount of flow consisting of agents experiencing disutility at most C' is as large as possible. Unfortunately, in general no such flow exists. The example can be found at Section 5.5 of Chapter 5 of this thesis.

Since the proofs for earliest arrival flows [Gale, 1959, Minieka, 1973, Wilkinson, 1971, Baumann and Skutella, 2006] show this stronger property which does not generalize, we take a different approach. Our proof is based on duality (of an infinite dimensional LP, though we do not require any technical results on such LPs). The main technical challenge in our work comes from determining the correct ansatz for the dual solution, as well

as exploiting properties of the residual networks obtained from the successive shortest paths algorithm in precisely the right way to demonstrate certain complementary slackness conditions.

We remark that some of the work on maximum flow over time does make the connection to infinite dimensional LPs; see [Sharkey, 2011] for a survey and some further references. In particular, we point out the flow-over-time version of max-flow min-cut by [Philpott, 1990], which can be viewed as a derivation of strong duality for the corresponding infinite linear program.

As was the case with the time-expanded graph approach, the optimal dual solution immediately provides us with corresponding tolls for which the optimal flow is an equilibrium. However, we obtain an explicit formula for the optimal tolls, in terms of the successive shortest paths of the graph (see Section 3.3). This may be useful in obtaining a better structural understanding of optimal tolls, beyond just their computation. We also remark that a corollary of our result is that there is always an optimal solution without waiting (except at the source).

Consider for a moment the model where users cannot choose their departure time, but instead are released from the source at a fixed rate ν_0, and simply wish to reach the destination as early as possible. This is the game-theoretic model that has received the most attention from the flow over time perspective (see Section 1.2 of Chapter 1 of this thesis). Our construction of optimal tolls is applicable to this model as well, as discussed in Section 3.5. As far as we are aware, no explicit description of optimal tolls was previously known even in this setting.

We now return to the subtlety alluded to earlier: the distinction between strongly enforcing an optimal flow, and only weakly enforcing it.

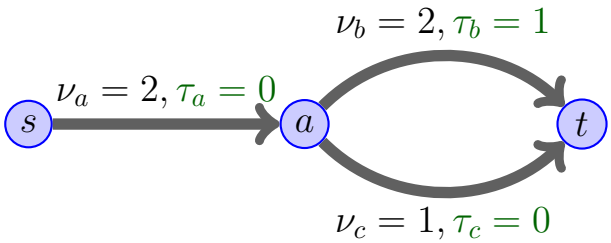


Figure 3.1: An instance where time-varying arc tolls cannot enforce that *all* equilibria are optimal flows.

Consider the simple instance in Figure 3.1. Suppose that the outflow of arc a is larger than 1 for some period in the optimum flow, due to the choice of scheduling cost function. In this period, one unit of flow would take the bottom arc c, and the rest will be routed on b. Since the total cost (including tolls) of all users is the same in a tolled dynamic equilibrium, a toll of cost equivalent to a unit delay on arc c is needed in this period to induce the optimal flow. But then it will also be an equilibrium to send *all* flow in this period along b.

To strongly enforce an optimal flow, we need more flexible tolls. One way that we can do it is by "tolling lanes". If we are allowed to dynamically divide up the capacity of an

arc into "lanes" (say a "fast lane" and a "slow lane"), and then separately set time-varying tolls on each lane, then we *can* strongly enforce any optimal flow. We discuss this further in Section 3.5. We are not aware of settings where this phenomenon has been previously observed, and it would be interesting to explore this further in a more applied context.

Outline of the chapter. We introduce some basic notation and notions, as well as formally define our model, in Section 3.2. In Section 3.3, we describe our algorithm, and show that it returns a feasible flow over time; we restrict ourselves to the most relevant case of a strictly unimodal scheduling cost function. In Section 3.4 we show optimality of this algorithm, and in Section 3.5 we derive optimal tolls from this analysis. Finally, in Section 3.6 we discuss general scheduling cost functions.

3.2 Model and preliminaries

We consider flow over time with waiting on the vertices. We refer to Chapter 2 for the network structure, notation and flow over time definitions.

We assume all graphs to be simple, and that there are no digons (i.e., there are no pairs $v, w \in V$ so that vw and wv are both arcs). This is for notational convenience only—this restriction can easily be lifted.

The notation $(z)^+$ is used to denote the nonnegative part of z, i.e., $(z)^+ := \max\{z, 0\}$. Given $v \in \mathbb{R}^X$ and $A \subseteq X$, we will use the shorthand notation $v(A) := \sum_{a \in A} v(a)$.

Residual networks. Given a static s-t-flow f, its *residual network* $G^f = (V, E^f)$ is defined by

$$E^f = \{vw : vw \in E \text{ and } f_{vw} < \nu_{vw}\} \cup \{vw : wv \in E \text{ and } f_{wv} > 0\}.$$

Call arcs in $E^f \cap E$ *forward arcs* and arcs in $E^f \setminus E$ *backwards arcs*. The *residual capacity* ν^f_e of an arc $e \in E^f$ is then $\nu^f_{vw} := \nu_{vw} - f_{vw}$ for vw a forward arc, and $\nu^f_{vw} := f_{wv}$ for vw a backwards arc. We also define $\tau_{vw} := -\tau_{wv}$ for all backwards arcs vw.

Given a subset $F \subseteq E$, we use $\chi(F)$ to denote the characteristic vector of F. In particular, if P is a path from v to w, then $\chi(P)$ is a unit flow from v to w.

We make the definitions $\overleftarrow{E} := \{wv : vw \in E\}$ and $\overleftrightarrow{E} := E \cup \overleftarrow{E}$. We will regard a vector $g \in \mathbb{R}^{\overleftrightarrow{E}}_+$ as a flow in $(V, \overleftrightarrow{E})$ if for every $vw \in E$, either $g_{vw} = 0$ or $g_{wv} = 0$. Given two such flows f and g, we define their sum $f + g$ by taking the sum as vectors, and then cancelling flows on oppositely directed arcs if necessary (so $(f + g)_{vw}$ and $(f + g)_{wv}$ are never both nonzero). Define $f - g$ similarly.

Given a choice of value Q, a *minimum cost flow* is an s-t-flow f^* minimizing $\sum_{e \in E} f_e \tau_e$ (amongst all s-t-flows f of value Q). An s-t-flow f (of the correct value) is a minimum cost flow if and only if E^f contains no negative cost cycles, i.e., cycles $C \subseteq E^f$ with $\tau(C) < 0$.

We also have a natural notion of a residual network in the flow over time setting. Given a flow over time f with waiting on the vertices, define, for any $\theta \in \mathbb{R}$,

$$E^f(\theta) = \{vw : vw \in E \wedge f_{vw}(\theta) < \nu_{vw}\} \cup \{vw : wv \in E \wedge f_{wv}(\theta - \tau_{wv}) > 0\}.$$

Minimizing scheduling cost. We are concerned with the following optimization problem. Given a scheduling cost function $\rho : \mathbb{R} \to \mathbb{R}_+$, as well as a value $\alpha > 0$, determine a flow over time f of value Q that minimizes the sum of the journey time cost $\alpha \sum_{e \in E} \tau_e \cdot \int_{-\infty}^{\infty} f_e(\theta) d\theta$ and the scheduling cost $\int_{-\infty}^{\infty} \nabla f_t(\theta) \cdot \rho(\theta) d\theta$. As already discussed, we assume that ρ satisfies the growth bound, i.e., that $\theta \to \rho(\theta) + \alpha\theta$ is nondecreasing. This ensures that waiting at t is not needed, which is in fact disallowed by our definition[1], and makes various arguments cleaner. We will also make the assumption that ρ is strongly unimodal[2]. We then assume w.l.o.g. that the minimizer of ρ is at 0, and that $\rho(0) = 0$. For further technical convenience, by adjusting ρ on a set of measure zero we take ρ to be lower semi-continuous[3]..

The above conditions will suffice for our structural characterization of an optimum flow and its analysis, but more is needed in order to be able to implement the algorithm. The algorithm will require not just oracle access to ρ, but also to ρ^{-1}. That is, given $y > 0$, we are able to query the pair of solutions (one positive, one negative) that map to y under ρ. In order to ensure that the optimal solution has a rational description, we should require not only that ρ maps rationals to rationals, but that ρ^{-1} does too; a simple function like $\rho(\theta) = \theta^2$ that violates this can lead to irrational optimum solutions, as we will remark on later. For algorithmic purposes, it is sensible to restrict attention to scheduling costs that are explicitly given as piecewise linear functions; we will focus primarily on this case.

The assumption of strong unimodality is not necessary; the algorithm and analysis can be extended (with some additional effort). We postpone this discussion to the end of the chapter.

3.3 A combinatorial algorithm

In this section we present an algorithm that computes an optimal flow over time, assuming that ρ is strongly unimodal. The proof of optimality is discussed in Section 3.4.

We begin by recalling the *successive shortest paths (SSP)* algorithm for computing a minimum cost static flow. It is not a polynomial time algorithm, so it is inefficient as an algorithm for static flows, but it provides a structure that is relevant for flows over time. This is of course well known from its role in constructing earliest arrival flows, which we will briefly detail.

The SSP algorithm construct a sequence of paths $(P_1, P_2, \ldots)$ and associated amounts $(x_1, x_2, \ldots)$ inductively as follows. Suppose $P_1, \ldots, P_j$ and $x_1, \ldots, x_j$ have been defined. Let

$$f^{(j)} = \sum_{i=1}^{j} x_i \chi(P_i),$$

and let G_j denote the residual graph of $f^{(j)}$ (G_0 being the original network). Also let $d_j(v, w)$ denote the length (w.r.t. arc free transit times τ in G_j) of a shortest path from v to w in G_j (this may be infinite). By construction, G_j will contain no negative cost cycles,

[1] Were this really needed, one could simply add a dummy arc tt' to a new sink t'.

[2] I.e., strictly decreasing until some moment, and then strictly increasing.

[3] Since an increasing function is continuous almost everywhere, we can replace $\rho(\theta)$ by $\lim_{\epsilon \downarrow 0} \rho(\theta + \epsilon)$ for all $\theta \geq 0$; and similarly with $\lim_{\epsilon \uparrow 0} \rho(\theta + \epsilon)$ for $\theta < 0$.

so that d_j is computable. If $d_j(s,t) = \infty$, we are done; set $m := j$. Otherwise, define P_{j+1} to be any shortest s-t-path in G_j, and x_{j+1} the minimum capacity in G_j of an arc in P_{j+1}. It can be shown that $\sum_{j=1}^{r} \tilde{x}_j \chi(P_j)$, with r and $\tilde{x}$ defined such that $\tilde{x}_j = x_j$ for $j < r$, $0 \leq \tilde{x}_r \leq x_r$ and $\sum_{j=1}^{r} \tilde{x}_j = M$, is a minimum cost flow of value M, as long as M is not larger than the value of a maximum flow.

To construct an earliest arrival flow with time horizon T, we (informally) send flow at rate x_j along path P_j for the time interval $[0, T - \tau(P_j)]$, for each $j \in [m]$ (if $\tau(P_j) > T$, we send no flow along the path). By this, we mean that for each $e = vw \in P_j$, we increase by x_j the value of $f_e(\theta)$ for $\theta \in [d_{j-1}(s,v), T - d_{j-1}(v,t)]$ (or if e is a backwards arc, we instead decrease $f_{wv}(\theta - \tau_{wv})$). An argument is needed to show that this defines a valid flow, since we must not violate the capacity constraints, and moreover, P_j may contain reverse arcs not present in G (see, e.g., [Skutella, 2009]).

We are now ready to describe our algorithm for minimizing the disutility, which is a natural variation on the earliest arrival flow algorithm. It is also constructed from the successive shortest paths, but using a *cost horizon* rather than a *time horizon*. For now, consider C to be a given value (it will be the "cost horizon"). For each $j \in [m]$ with $\alpha d_{j-1}(s,t) \leq C$, we send flow at rate x_j along path P_j for the time interval $[a_j, b_j]$ chosen maximally so that

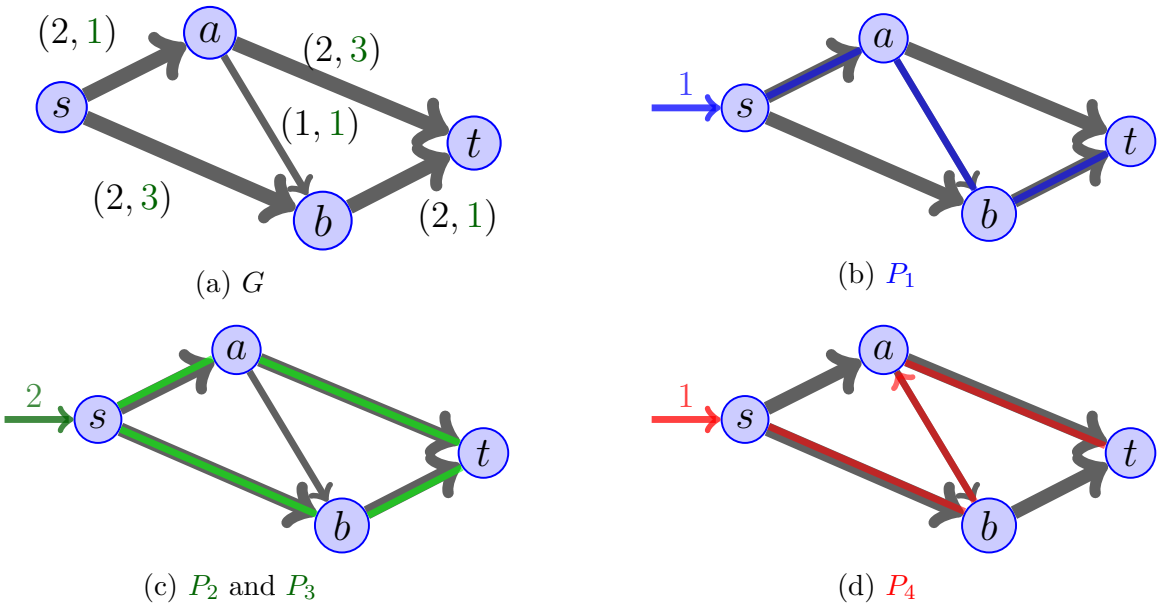
$$\rho(\xi + d_{j-1}(s,t)) \leq C - \alpha d_{j-1}(s,t) \quad \text{for all} \quad \xi \in [a_j, b_j].$$

(If ρ is continuous, then of course $\rho(a_j + d_{j-1}(s,t)) = \rho(b_j + d_{j-1}(s,t)) = C - \alpha d_{j-1}(s,t)$). Note that a user leaving at time a_j or b_j and using path P_j, without waiting at any moment, incurs disutility C; whereas a user leaving at some time $\theta \in (a_j, b_j)$ and using path P_j will incur a strictly smaller travel cost.

As we will shortly argue, this results in a feasible flow over time f. Given this, its value will be $\sum_{j=1}^{m} x_j(b_j - a_j)$. Since ρ is strongly unimodal, this value changes continuously and monotonically with C. Thus a bisection search can be used to determine the correct choice of C for a given value Q, at least to within some predetermined error ϵ. Determining the *precisely* correct value of C may not be possible without some additional information about ρ.

If ρ is piecewise linear (as is the case, in particular, for the "standard" β/γ choice generally used in the transportation economics literature), bisection search can be avoided. Let K_l and K_r denote the number of linear segments of ρ to the left and right of 0, respectively, and let $K = K_l + K_r$. Write $a_j(C)$ and $b_j(C)$ to explicitly indicate the dependence of the interval in which flow is sent along P_j as a function of C. Then let $Q_j(C) := x_j(b_j(C) - a_j(C))$; this is the total mass sent along path P_j in the solution obtained with time horizon C. Now notice that $a_j(C)$ is a piecewise linear function with at most $K_l + 1$ linear segments; it is defined by $\rho(a_j(C) + d_{j-1}(s,t)) = C - \alpha d_{j-1}(s,t)$ for $C \geq \alpha d_{j-1}(s,t)$, and $a_j(C) = -d_{j-1}(s,t)$ otherwise. Similarly, $b_j(C)$ is a piecewise linear function with at most $K_r + 1$ linear segments. The total value $\sum_{j=1}^{m} Q_j(C)$ is thus piecewise linear with at most $m(K+2)$ linear segments. Thus, even the entire parametric curve of cost horizon C against flow value Q can be computed in time $O(mK)$, once the successive shortest paths have been computed.

Before proving the correctness of our algorithm, we show an example of a flow over time minimizing a given scheduling cost, as would be constructed by our algorithm.



(a) G (b) P_1

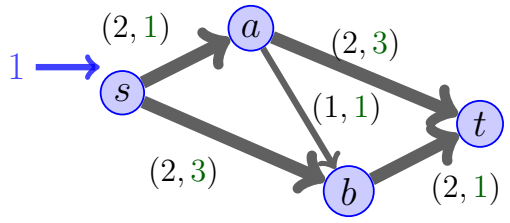
(c) P_2 and P_3 (d) P_4

Figure 3.2: The network and the successive shortest paths of Example 3.1.

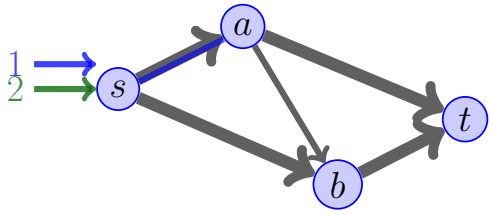
Example 3.1. Consider the graph $G = (V, E)$ illustrated in Figure 3.2a, with $V = \{s, a, b, t\}$, $E = \{sa, sb, ab, at, bt\}$ and capacities ν_e and free transit times τ_e as indicated in the figure, in the order (ν_e, τ_e).

The successive shortest paths are $P_1 = \{s, a, b, t\}$, with length 3, $P_2 = \{s, a, t\}$ and $P_3 = \{s, b, t\}$, both with length 4, and $P_4 = \{s, b, a, t\}$ with length 5. All the associated amounts are equal to 1 (see Figures 3.2b to 3.2d).

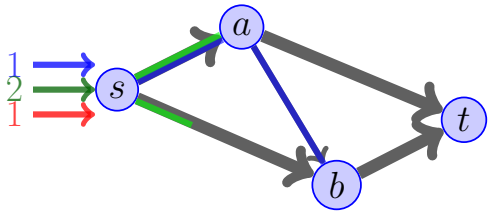
Consider now a cost horizon C equal to 6 and the standard scheduling cost function given in Equation (1.1) with $\alpha = 1$, $\beta = 0.5$, $\gamma = 2$ and $T^* = 0$. Our algorithm then sends 1 unit of flow along P_1 for the time interval $[-9, -1.5]$; 1 unit of flow along P_2 and 1 along P_3 for the time interval $[-8, -3]$; and 1 unit of flow along P_4 for the time interval $[-7, -4.5]$. The resulting flow is described in Figure 3.3 with a sequence of snapshots of the network.



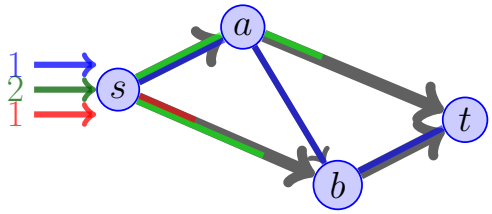
(a) Time −9



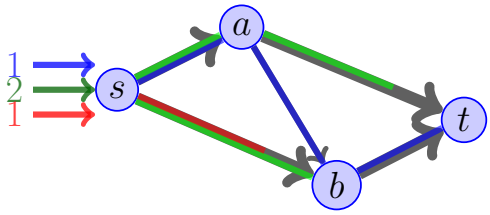
(b) Time −8



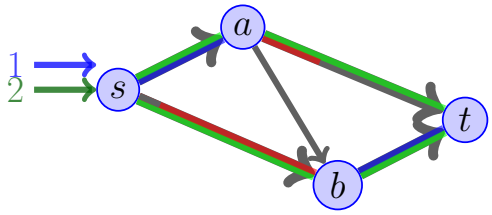
(c) Time −7



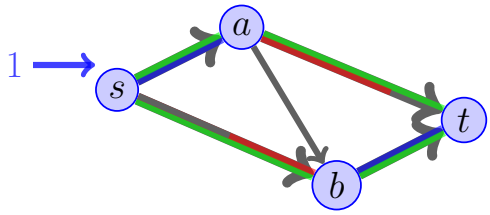
(d) Time −6



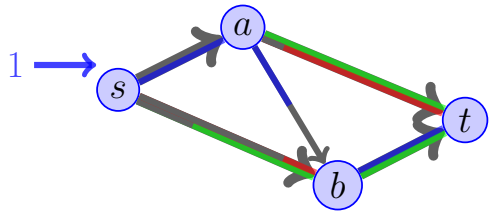
(e) Time −5



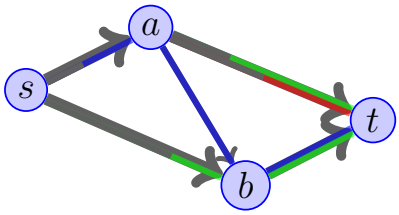
(f) Time −4



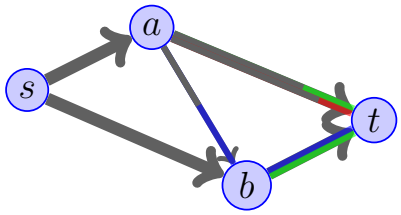
(g) Time −3



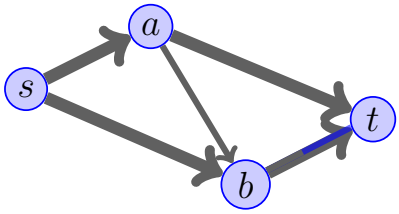
(h) Time −2



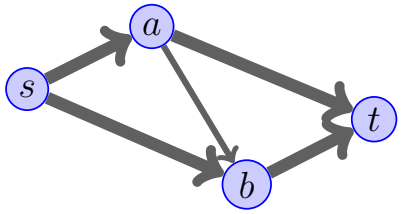
(i) Time −1



(j) Time 0



(k) Time 1



(l) Time 1.5

Figure 3.3: Chronological sequence of snapshots of the optimal flow for the instance given in Example 3.1.

Feasibility. In the following we show that the resulting flow f is a feasible flow over time. Given a vertex $v \in V$, a time $\theta \in \mathbb{R}$ and $j \in [m]$, let

$$c_j(v,\theta) = \alpha d_{j-1}(s,t) + \rho(\theta + d_{j-1}(v,t)).$$

If $v \in P_j$ then $c_j(v,\theta)$ is the travel cost of a user that utilizes path P_j and passes through vertex v at time θ; there does not seem to be a simple interpretation if $v \notin P_j$ however. Now define

$$J(v,\theta) = \max\{j \in [m] : c_j(v,\theta) \leq C\}, \tag{3.2}$$

with the convention that the maximum over the empty set is 0. We remark for future reference that since $d_m(s,t) = \infty$, we do have that

$$\alpha d_{J(v,\theta)}(s,t) + \rho(\theta + d_{J(v,\theta)}(v,t)) > C. \tag{3.3}$$

The motivation for this definition comes from the following theorem, which completely characterizes f in terms of the static flows arising from successive shortest paths. (If preferred, one could even think of this theorem as providing the definition of f.)

Theorem 3.1. *$f_{vw}(\theta) = f_{vw}^{(J(v,\theta))}$ for any $vw \in E$ and $\theta \in \mathbb{R}$.*

Before proving Theorem 3.1, we need the following lemma.

Lemma 3.2. *$c_j(v,\theta)$ is nondecreasing with j for any $\theta \in \mathbb{R}$.*

Proof. Consider any $j \in [m-1]$; we show that $c_{j+1}(v,\theta) \geq c_j(v,\theta)$. Suppose R is a shortest v-t-path in G_{j-1}, so $\tau(R) = d_{j-1}(v,t)$. Consider the unit v-t flow $g = \chi(P_{j+1}) - \chi(P_j) + \chi(R)$ in $\overleftrightarrow{E}$. Now observe that the support of g is contained in G_j: P_{j+1} and $\overleftarrow{P_j}$ are certainly contained in G_j; and if $e \in R \cap (E_{j-1} \setminus E_j)$, then $e \in P_j$, which means $g_e = 0$. Since G_j contains no negative cost cycles, the cost of g is at least that of a shortest v-t-path in G_j, and so

$$d_j(v,t) \leq \tau(P_{j+1}) - \tau(P_j) + \tau(R) = d_j(s,t) - d_{j-1}(s,t) + d_{j-1}(v,t)\ .$$

Finally, we can conclude

$$\begin{aligned}\alpha d_j(s,t) + \rho(\theta + d_j(v,t)) &= \alpha d_j(s,t) + \rho(\theta + d_{j-1}(v,t)) - \rho(\theta + d_{j-1}(v,t)) + \rho(\theta + d_j(v,t)) \\ &\geq \alpha d_j(s,t) + \rho(\theta + d_{j-1}(v,t)) - \alpha(d_j(v,t) - d_{j-1}(v,t)) \\ &\geq \alpha d_{j-1}(s,t) + \rho(\theta + d_{j-1}(v,t)),\end{aligned}$$

where the first inequality follows from the growth bound, using $d_j(v,t) \geq d_{j-1}(v,t)$. □

Proof of Theorem 3.1. Fix some $vw \in E$ and $\theta \in \mathbb{R}$. Consider now any $j \in [m]$ for which $\alpha\tau(P_j) \leq C$ (so that P_j is used for a nontrivial interval) and $vw \in P_j$. Since P_j is a shortest path in G_{j-1}, if we send flow along this path starting from some time ξ, it will arrive at v at time $\xi + d_{j-1}(s,v)$. Considering the definition of the interval $[a_j, b_j]$, we see that P_j contributes flow to vw at time θ if $c_j(v,\theta) \leq C$. By Lemma 3.2, this occurs precisely if $j \leq J(v,\theta)$.

Considering in similar fashion paths P_j with $wv \in P_j$ (and noting that $J(w, \theta + \tau_{vw}) = J(v,\theta)$), we determine that

$$f_{vw}(\theta) = \sum_{\substack{j: vw \in P_j \\ j \leq J(v,\theta)}} x_j - \sum_{\substack{j: wv \in P_j \\ j \leq J(v,\theta)}} x_j = f_{vw}^{(J(v,\theta))}.$$ □

Feasibility of f is now immediate.

Corollary 3.3. *f is a feasible flow over time without waiting.*

Proof. By the way that we constructed f, it has value Q, satisfies flow conservation, and has no waiting. Only nonnegativity and the capacity constraint remain, which follows from Theorem 3.1. □

3.4 Optimality

In this section, we show that our proposed algorithm does return an optimal flow.

3.4.1 Duality-based certificates of optimality

We can write the problem we are interested in as an infinite continuous linear program as follows:

$$\begin{aligned}
\min \quad & \int_{-\infty}^{\infty} \rho(\theta)\nabla f_t(\theta)d\theta + \alpha \sum_{e\in E} \tau_e \int_{-\infty}^{\infty} f_e(\theta)d\theta + \alpha \sum_{v\in V\setminus\{s,t\}} \int_{-\infty}^{\infty} z_v(\theta)d\theta \\
\text{s.t.} \quad & \int_{-\infty}^{\infty} \nabla f_s(\theta)d\theta = -Q \\
& \int_{-\infty}^{\infty} \nabla f_t(\theta)d\theta = Q \\
& \int_{-\infty}^{\theta} \nabla f_v(\xi)d\xi = z_v(\theta) \qquad \forall v \in V \setminus \{s,t\}, \theta \in \mathbb{R} \\
& f_e(\theta) \le \nu_e \qquad \forall e \in E, \theta \in \mathbb{R} \\
& z, f \ge 0
\end{aligned} \tag{3.4}$$

Here, $z_v(\theta)$ represents the amount of flow waiting at node v at time θ (which must always be nonnegative). Both f_e for any $e \in E$ and z_v for any $v \in V$ should be bounded and measurable functions with compact support. This implies that in fact z_v is absolutely continuous for each $v \in V$. Note that the objective function captures separately the contribution to the journey time coming from actually travelling across arcs, and the contribution from waiting at nodes. As a further remark, this linear program does not explicitly prevent flow from departing and then returning to t; only the aggregate constraint $\int_{-\infty}^{\infty} \nabla f_t(\theta) = Q$ is imposed. The growth condition on ρ, however, ensures that it is never profitable to do this; the reduction of scheduling cost is never more than the journey time cost (including waiting at nodes).

Given that this is an infinite-dimensional linear program, one may reasonably expect to be able to write down a dual, and make use of weak and strong duality, as well as complementary slackness conditions. However, care is needed: the situation for infinite (even countable) dimensional linear programs is subtle. Strong duality and even *weak* duality may fail to hold, even for infinite linear programs with a countable number of variables and constraints [Romeijn et al., 1992]. Here, our primal variables live in the space of bounded measurable functions, and there are an uncountably infinite set of constraints:

it is a *continuous* linear program (see [Sharkey, 2011] for a review of some of the relevant literature). Fortunately, the particular structure of our continuous linear program is of a well-behaved form.

A continuous time linear program [Reiland, 1980] is (after a possible change of variables) of the form

$$\begin{aligned} \min \quad & \int_{\tilde{T}_0}^{\tilde{T}_1} \tilde{c}^\top(\theta) y(\theta) d\theta \\ \text{s.t.} \quad & \tilde{B}(\theta) y(\theta) \geq \tilde{b}(\theta) + \int_{\tilde{T}_0}^{\theta} \tilde{K}(\xi, \theta) y(\xi) d\xi \qquad \forall\, \tilde{T}_0 \leq \theta \leq \tilde{T}_1 \\ & y(\theta) \geq 0 \qquad \forall\, \tilde{T}_0 \leq \theta \leq \tilde{T}_1. \end{aligned}$$

Here, $[\tilde{T}_0, \tilde{T}_1]$ is a compact interval, $\tilde{c}$ and $\tilde{b}$ are vectors of bounded measurable functions, and $\tilde{B}$, $\tilde{K}$ are matrices of bounded measurable functions. Note that $\tilde{K}(\xi, \theta)$ is only required for $\xi \leq \theta$. The components of a feasible solution y are required also to be bounded and measurable. The corresponding dual program is

$$\begin{aligned} \max \quad & \int_{\tilde{T}_0}^{\tilde{T}_1} \tilde{b}^\top(\theta) w(\theta) d\theta \\ \text{s.t.} \quad & \tilde{B}^\top(\theta) w(\theta) \leq \tilde{c}(\theta) + \int_{\theta}^{\tilde{T}_1} \tilde{K}^\top(\theta, \xi) w(\xi) d\xi \qquad \forall\, \tilde{T}_0 \leq \theta \leq \tilde{T}_1 \\ & w(\theta) \geq 0 \qquad \forall\, \tilde{T}_0 \leq \theta \leq \tilde{T}_1. \end{aligned}$$

Strong duality does not hold without further assumptions, but weak duality (and hence sufficiency of related complementary slackness conditions) do hold [Reiland, 1980, Theorem 1]. As a consequence, if solutions y and w are feasible to the primal and dual, and satisfy

$$\begin{aligned} \int_{\tilde{T}_0}^{\tilde{T}_1} \left[w^\top(\theta) \left(\tilde{b}(\theta) - \int_{\tilde{T}_0}^{\theta} \tilde{K}(\xi, \theta) y(\xi) d\xi - \tilde{B}(\theta) y(\theta) \right) \right] = 0, \\ \int_{\tilde{T}_0}^{\tilde{T}_1} \left[y^\top(\theta) \left(\tilde{c}(\theta) - \int_{\theta}^{\tilde{T}_1} \tilde{K}(\theta, \xi) w(\xi) d\xi - \tilde{B}^\top(\theta) w(\theta) \right) \right] = 0, \end{aligned}$$

then y and w are both optimal. Reiland [1980] gives constraint qualifications under which a version of this is both necessary and sufficient for optimality, i.e., where strong duality holds; we will not need this, and so don't discuss this further here.

Our continuous LP (3.4) fits within this class. First, we note that while we wrote the program with an unbounded interval, this was purely for notational convenience; any optimal solution must be contained in the interval $\{\theta : |\theta| \leq Q/\nu_{\text{SP}} + \tau_{\text{SP}}\}$, where ν_{SP} is the minimum capacity of an arc of some shortest s-t-path in G, and τ_{SP} is the length of this path. (This comes from considering a solution that minimizes the average journey time, and has minimum scheduling cost subject to this.) One may introduce the additional variables $F_e(\theta) = \int_{-\infty}^{\theta} f_e(\xi) d\xi$, after which it is straightforward to place things in the desired form.

After writing down the dual constraints and the complementary slackness conditions, the following sufficient conditions for optimality to (3.4) are obtained. The theorem is stated only for the case where the primal flow has no waiting: our algorithm produces

such a flow, and so this is the case of interest to us (it will thus be an immediate corollary of our result that there is always an optimal flow without waiting). For completeness and convenience, we give a short, self-contained proof of this theorem; none of the above discussion will be used.

Theorem 3.4. *Let f be a flow over time without waiting and with value Q, and suppose that $\pi : V \times \mathbb{R} \to \mathbb{R}$ satisfies the following, for some choice of C:*

(i) $\theta \to \pi_v(\theta) - \alpha\theta$ *is nonincreasing.*

(ii) $\pi_w(\theta + \tau_{vw}) \leq \pi_v(\theta) + \alpha\tau_{vw}$ *for all* $\theta \in \mathbb{R}, vw \in E^f(\theta)$.

(iii) $\pi_s(\theta) = 0$ *for all* $\theta \in \mathbb{R}$.

(iv) $\pi_t(\theta) = (C - \rho(\theta))^+$ *for all* $\theta \in \mathbb{R}$, *and* $\nabla f_t(\theta) = 0$ *whenever* $\rho(\theta) > C$.

Then f is an optimal solution.

Proof. We will need the following technical lemma (obvious via integrating by parts in the case that h is also absolutely continuous).

Claim 3.5. *Let $h : \mathbb{R} \to \mathbb{R}$ be a nonincreasing function, and $z : \mathbb{R} \to \mathbb{R}_+$ be an absolutely continuous nonnegative function with compact support. Then $\int_{-\infty}^{\infty} h(\theta)z'(\theta)d\theta \geq 0$.*

Proof. Since $\int_{-\infty}^{\infty} z'(\theta)d\theta = 0$, we may assume by shifting if necessary that h is nonnegative on the support of z. Let μ be a measure so that $\mu([\theta, \infty)) = h(\theta)$ for almost every θ in the support of z. We certainly have that for any θ,

$$\int_{-\infty}^{\theta} z'(\xi)d\xi = \Big[z(\xi)\Big]_{-\infty}^{\theta} = z(\theta) - 0 \geq 0.$$

Thus

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}_{\xi\leq\theta} z'(\xi)d\xi d\mu(\theta) \geq 0,$$

from which we obtain the result by applying Fubini's theorem. □

Define, for each $vw \in E$,

$$\mu_{vw}(\theta) := (\pi_w(\theta + \tau_{vw}) - \pi_v(\theta) - \alpha\tau_{vw})^+.$$

Now let g, z be any feasible solution to (3.4) with compact support. Consider any $v \in V \setminus \{s, t\}$, and observe that

$$\begin{aligned}\int_{-\infty}^{\infty} \pi_v(\theta)\nabla g_v(\theta) + \alpha z_v(\theta)d\theta &= \int_{-\infty}^{\infty} (\pi_v(\theta) - \alpha\theta)\nabla g_v(\theta)d\theta \;+\; \alpha\int_{-\infty}^{\infty} \theta\nabla g_v(\theta) + z_v(\theta)d\theta \\ &= \int_{-\infty}^{\infty} (\pi_v(\theta) - \alpha\theta)\nabla g_v(\theta)d\theta \;+\; \Big[\alpha\theta z_v(\theta)\Big]_{-\infty}^{\infty} \\ &\geq 0.\end{aligned} \tag{3.5}$$

The final inequality comes from observing that the first term is nonnegative by Claim 3.5 (applied with $h(\theta) = \pi_v(\theta) - \alpha\theta$, which is nonincreasing by property (i), and $z(\theta) = z_v(\theta)$), and that the second term is zero since z has compact support.

We then have the following sequence of inequalities (detailed explanations for each step follow).

$$\begin{aligned}
\text{cost}(g) &= \int_{-\infty}^{\infty} \rho(\theta)\nabla g_t(\theta)d\theta + \sum_{e\in E}\int_{-\infty}^{\infty} \alpha\tau_e g_e(\theta)d\theta + \alpha \sum_{v\in V\setminus\{s,t\}}\int_{-\infty}^{\infty} z_v(\theta)d\theta \\
&\overset{(*)}{\geq} \int_{-\infty}^{\infty} (C - \pi_t(\theta))\nabla g_t(\theta)d\theta + \sum_{vw\in E}\int_{-\infty}^{\infty} \Big(\pi_w(\theta+\tau_{vw}) - \pi_v(\theta) - \mu_{vw}(\theta)\Big) g_{vw}(\theta)d\theta \\
&\quad + \alpha \sum_{v\in V\setminus\{s,t\}}\int_{-\infty}^{\infty} z_v(\theta)d\theta \\
&\overset{(**)}{=} CQ + \sum_{v\in V\setminus\{s,t\}}\int_{-\infty}^{\infty} [\pi_v(\theta)\nabla g_v(\theta) + \alpha z_v(\theta)]d\theta - \sum_{e\in E}\int_{-\infty}^{\infty} \mu_e(\theta)g_e(\theta)d\theta \\
&\overset{(***)}{\geq} CQ - \sum_{e\in E}\int_{-\infty}^{\infty} \mu_e(\theta)\nu_e d\theta.
\end{aligned}$$

Inequality $(*)$ follows from property (iv) of π, along with the definition of μ_e. The equality $(**)$ follows by recombining the $g_e(\theta)$ terms and recalling that $\pi_s \equiv 0$ and that g has value Q. Finally, $(***)$ follows from (3.5), and the inequalities $\mu_e(\theta) \geq 0$ and $g_e(\theta) \leq \nu_e$ that hold for all $e \in E$ and $\theta \in \mathbb{R}$.

To complete the proof of the theorem, we now observe that all of the inequalities in the above hold with equality if $g = f$ and (consistent with the no-waiting assumption on f) $z = 0$. Property (ii) implies that if $f_{vw}(\theta) > 0$ (so that $wv \in E^f(\theta)$), then $\mu_{vw}(\theta) = \pi_w(\theta + \tau_{vw}) - \pi_v(\theta) - \alpha\tau_{vw}$, yielding equality in $(*)$. It also implies that if $f_{vw}(\theta) < \nu_{vw}$ (so that $vw \in E^f(\theta)$) then $\mu_{vw}(\theta) = 0$. This, together with $z = 0$, implies the equality in $(***)$. □

As is often the case, the optimal dual solution also provides us the prescription for tolls to induce the optimum flow. We delay this discussion to Section 3.5.

3.4.2 The dual prescription

We now give a certificate of optimality $\pi : V \times \mathbb{R} \to \mathbb{R}$ for (3.4) that satisfies the conditions of Theorem 3.4. Given a vertex $v \in V$ and a time $\theta \in \mathbb{R}$ let

$$\pi_v(\theta) = \max\{\hat{\pi}_v(\theta), \bar{\pi}_v(\theta), 0\}$$

where

$$\begin{aligned}
\hat{\pi}_v(\theta) &= -\alpha d_{J(v,\theta)}(v, s), \\
\bar{\pi}_v(\theta) &= C - \alpha d_{J(v,\theta)}(v, t) - \rho(\theta + d_{J(v,\theta)}(v, t)).
\end{aligned}$$

Some intuition for this choice of π can be obtained by thinking in terms of "temporal" shortest paths in the residual $E^f(\theta)$ of the flow f returned by the algorithm. For some

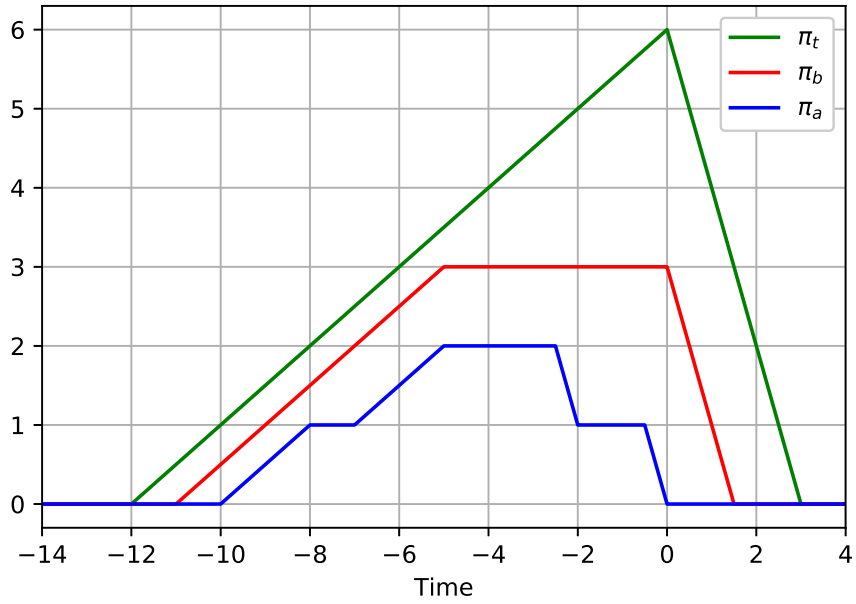


Figure 3.4: Dual values to show optimality of the flow in Example 3.1. $\pi_s(\theta) = 0$ for all $\theta \in \mathbb{R}$, and is not shown.

$v \in V$ and $\theta \in \mathbb{R}$, consider a shortest s-v-path $P = (s = v_0, v_1, \ldots, v_{k-1}, v_k = v)$ in $f^{(J(v,\theta))}$. This path can be turned into a "temporal" path that ends at v at time θ, in the obvious way: the path should visit node v_i at time θ_i, such that $\theta_i = \theta_{i-1} + \tau_{v_{i-1}v_i}$ for each $i \in [k]$, and $\theta_k = \theta$. It turns out that every arc in this temporal path lies in the residual $E^f(\theta)$, and given that $\pi_s \equiv 0$, this implies an upper bound on $\pi_v(\theta)$ by Theorem 3.4 property (ii); this upper bound motivates the definition of $\hat{\pi}$. A similar consideration of a shortest t-v-path in $G^{(J(v,\theta))}$, along with the requirement that $\pi_t(\theta) = (C - \rho(\theta))^+$, motivates $\bar{\pi}$.

Example 3.2. Figure 3.4 shows the dual values π_v for the instance of Example 3.1. The conditions for Theorem 3.4 are all satisfied; for example, $0 < f_{sb}(-7) < \nu_{sb}$, and so we should have that $\pi_b(-7+3) = \pi_s(-7) + \alpha \cdot 3 = 3$, which is indeed the case.

Lemma 3.6. *We have $\pi_s(\theta) = 0$ and $\pi_t(\theta) = (C - \rho(\theta))^+$ for all $\theta \in \mathbb{R}$.*

Proof. Notice that

$$\begin{aligned} & \hat{\pi}_s(\theta) = -\alpha d_{J(s,\theta)}(s,s) = 0 \\ \text{and} \qquad & \bar{\pi}_s(\theta) = C - \alpha d_{J(s,\theta)}(s,t) - \rho(\theta + d_{J(s,\theta)}(s,t) < 0; \end{aligned}$$

the last inequality comes from (3.3). Thus $\pi_s(\theta) = 0$ for all θ.

Next, set $j = J(t,\theta)$ and observe that

$$\bar{\pi}_t(\theta) = C - \alpha d_j(t,t) - \rho(\theta + d_j(t,t)) = C - \rho(\theta).$$

For $\hat{\pi}_t(\theta)$, we consider two cases.

- If $j = 0$, then $\hat{\pi}_t(\theta) = -\alpha d_0(t, s) \leq 0$.
- If $j \geq 1$, then by (3.2) (and using $d_{j-1}(t, t) = 0$),

$$\alpha d_{j-1}(s, t) + \rho(\theta) \leq C.$$

  Since $d_{j-1}(s, t)$ is equal to the length of path P_j, and the reverse of P_j is a t-s-path in G_j, we deduce that $d_j(t, s) \leq -d_{j-1}(t, s)$, and hence that $\hat{\pi}_t(\theta) = \alpha d_j(t, s) \leq C - \rho(\theta)$.

In either case, $\pi_t(\theta) = \max\{C - \rho(\theta), 0\}$. □

Thus conditions (iii) and (iv) of Theorem 3.4 hold. For the remaining conditions, we begin with some basic facts about distance labels associated with successive shortest paths (statements of a similar flavour can be found in Ahuja et al. [1993], for example).

Lemma 3.7. *For every $v \in V$, $d_j(v, s)$ is nonincreasing with j, and $d_j(v, t)$ is nondecreasing in j.*

Proof. We show that $d_{j-1}(v, s) \geq d_j(v, s)$ for all $v \in V$ and $j \in [m]$. If $d_{j-1}(v, s)$ is infinite, there is nothing to prove; by possibly restricting to a subgraph in the following argument, assume that all nodes are reachable from s in G_{j-1}. For any node labels $\sigma \in \mathbb{R}^V$, and any $vw \in \overleftrightarrow{E}$, let $c^\sigma_{vw} := c_{vw} + \sigma_w - \sigma_v$. Notice that for any ℓ, if σ is such that $c^\sigma_{vw} \geq 0$ for all $vw \in E_\ell$, then $\sigma_v - \sigma_s \geq d_\ell(s, v)$ for all $v \in V$.

Now define $\sigma_v = d_{j-1}(s, v)$ for each node v; then $c^\sigma_{vw} \geq 0$ for all $vw \in E_{j-1}$, with equality if v lies on a shortest path from s to w. Thus all arcs vw in the path P_j satisfy $c^\sigma_{vw} = 0$. Hence $c^\sigma_{vw} \geq 0$ for all $vw \in E_j$, and hence $d_{j-1}(s, v) = \sigma_v - \sigma_s \geq d_j(s, v)$ for all $v \in V$.

A similar argument "in reverse" applies for distances to t. This time, define $\sigma_v = d_j(v, t)$; then $c^\sigma_{vw} \geq 0$ for all $vw \in E_j$, with equality for all arcs of P_j. So $c^\sigma_{vw} \geq 0$ for all $vw \in E_{j-1}$, and hence $d_j(v, t) = \sigma_t - \sigma_v \geq d_{j-1}(v, t)$ for all nodes v. □

Lemma 3.8. *For all $j \in [m]$ and $v \in V$, $d_{j-1}(v, t) - d_{j-1}(s, t) = d_j(v, s)$.*

Proof. First of all, notice that $d_{j-1}(v, t)$ is finite precisely if $d_j(v, s)$ is, since a v-t-path in G_{j-1} can be combined with the reverse of P_j to obtain a v-s-path in G_j, and vice versa. Since $d_{j-1}(s, t)$ is always finite, the claim holds if $d_{j-1}(v, t) = d_j(v, s) = \infty$, so we assume both are finite in what follows.

Let R be a shortest v-s-path in G_j, and let R' be a v-t-path contained in $P_j + R$ (arcs in opposite directions are cancelled). Then R' is in G_{j-1}; if e is an arc in R not in G_{j-1}, then e is the reverse of an arc of P_j, and hence not in $P_j + R$. So $d_{j-1}(v, t) \leq \tau(R')$. But since G_{j-1} has no negative cost cycles, $\tau(R') \leq \tau(R) + \tau(P_j)$.

To show that $d_{j-1}(v, t) - d_{j-1}(s, t) \geq d_j(v, s)$, let $\bar{R}$ be a shortest v-t-path in G_{j-1}. Let w be the first (i.e., closest to v) vertex present in both $\bar{R}$ and P_j (notice that w might be equal to v or t) and let R be the v-w-path contained in $\bar{R}$. Then

$$d_{j-1}(v, t) = d_{j-1}(w, t) + \tau(R). \tag{3.6}$$

Since $R \subseteq E_j$, we have that:

$$\begin{aligned} d_j(v,s) &\leq d_j(w,s) + \tau(R) \\ &= d_{j-1}(w,t) - d_{j-1}(s,t) + \tau(R) && \text{since } w \in P_j \\ &= d_{j-1}(v,t) - d_{j-1}(s,t) && \text{by (3.6).} \end{aligned}$$

This concludes the proof. □

Now we are ready to show that π satisfies conditions (i) and (ii) of Theorem 3.4.

Lemma 3.9. *$\theta \to \pi_v(\theta) - \alpha\theta$ is nonincreasing.*

Proof. Fix any $\theta \in \mathbb{R}$ and $\epsilon \geq 0$. We show that $\pi_v(\theta) \geq \pi_v(\theta+\epsilon) - \alpha\epsilon$. Let $j := J(v,\theta)$ and $\ell := J(v,\theta+\epsilon)$.

- **Case 1:** $\pi_v(\theta+\epsilon) = -\alpha d_\ell(v,s)$.

  If $\ell \leq j$, then by Lemma 3.7

  $$\pi_v(\theta) \geq \hat{\pi}_v(\theta) = -\alpha d_j(v,s) \geq -\alpha d_\ell(v,s) = \pi_v(\theta+\epsilon).$$

  So suppose $\ell > j$. By the definition of $J(v,\theta+\epsilon)$, we know that

  $$\alpha d_{\ell-1}(s,t) + \rho(\theta + \epsilon + d_{\ell-1}(v,t)) \leq C. \tag{3.7}$$

  As a consequence, we have that:

  $$\begin{aligned} \pi_v(\theta) &\geq \bar{\pi}_v(\theta) \\ &= C - \alpha d_j(v,t) - \rho(\theta + d_j(v,t)) \\ &\overset{(*)}{\geq} C - \alpha d_j(v,t) - \rho(\theta + \epsilon + d_{\ell-1}(v,t)) - \alpha\left(\epsilon + d_{\ell-1}(v,t) - d_j(v,t)\right) \\ &\geq \alpha d_{\ell-1}(s,t) - \alpha\epsilon - \alpha d_{\ell-1}(v,t) && \text{by (3.7)} \\ &= -\alpha\epsilon - \alpha d_\ell(v,s) && \text{by Lemma 3.8} \\ &= \pi_v(\theta+\epsilon) - \alpha\epsilon. \end{aligned}$$

  Inequality $(*)$ follows from the growth bound on ρ combined with the fact that $\theta + \epsilon + d_{\ell-1}(v,t) \geq \theta + d_j(v,t)$ by Lemma 3.7.

- **Case 2:** $\pi_v(\theta+\epsilon) = C - \alpha d_\ell(v,t) - \rho(\theta + \epsilon + d_\ell(v,t))$.

  If $\ell \geq j$, then:

  $$\begin{aligned} \pi_v(\theta) &\geq \bar{\pi}_v(\theta) \\ &= C - \alpha d_j(v,t) - \rho(\theta + d_j(v,t)) \\ &= C - \alpha d_j(v,t) - \rho(\theta + \epsilon + d_\ell(v,t)) + \rho(\theta + \epsilon + d_\ell(v,t)) - \rho(\theta + d_j(v,t)) \\ &\geq C - \alpha d_j(v,t) - \rho(\theta + \epsilon + d_\ell(v,t)) - \alpha\left(\epsilon + d_\ell(v,t) - d_j(v,t)\right) \\ &= C - \rho(\theta + \epsilon + d_\ell(v,t)) - \alpha\epsilon - \alpha d_\ell(v,t) \\ &= \pi_v(\theta+\epsilon) - \alpha\epsilon. \end{aligned}$$

The second inequality follows again from the growth bound, this time combined with the inequality $d_\ell(v,t) \geq d_j(v,t)$.

If $\ell < j$, by definition of $J(v, \theta+\epsilon)$ we have that

$$\alpha d_\ell(s,t) + \rho(\theta + \epsilon + d_\ell(v,t)) > C.$$

From this, we obtain

$$\begin{aligned}
\pi_v(\theta) &\geq \hat{\pi}_v(\theta) \\
&= -\alpha d_j(v,s) \\
&> C - \alpha d_\ell(s,t) - \rho(\theta+\epsilon+d_\ell(v,t)) - \alpha d_j(v,s) \\
&\geq C - \alpha d_\ell(s,t) - \rho(\theta+\epsilon+d_\ell(v,t)) - \alpha d_{\ell+1}(v,s) && \text{by Lemma 3.7} \\
&= C - \alpha d_\ell(v,t) - \rho(\theta+\epsilon+d_\ell(v,t)) && \text{by Lemma 3.8} \\
&= \pi_v(\theta+\epsilon).
\end{aligned}$$

- **Case 3:** $\pi_v(\theta+\epsilon) = 0$.

  This case is immediate from the definition of π_v. □

Lemma 3.10. *If $vw \in E^f(\theta)$, then $\pi_w(\theta+\tau_{vw}) \leq \pi_v(\theta) + \alpha\tau_{vw}$.*

Proof. Let $j := J(v,\theta)$ and $\ell := J(w, \theta+\tau_{vw})$. Note that since $vw \in E^f(\theta)$, Theorem 3.1 implies that $vw \in E_j$.

- **Case 1:** $\pi_w(\theta+\tau_{vw}) = -\alpha d_\ell(w,s)$.

  If $\ell \leq j$, then

  $$\begin{aligned}
  \pi_v(\theta) &\geq -\alpha d_j(v,s) \\
  &\geq -\alpha\tau_{vw} - \alpha d_j(w,s) && \text{since } vw \in E_j \\
  &\geq -\alpha\tau_{vw} - \alpha d_\ell(w,s) && \text{by Lemma 3.7} \\
  &= \pi_w(\theta+\tau_{vw}) - \alpha\tau_{vw}.
  \end{aligned}$$

  So suppose $\ell > j$. By the definition of $J(w, \theta+\tau_{vw})$ we know that

  $$\alpha d_{\ell-1}(s,t) + \rho(\theta + \tau_{vw} + d_{\ell-1}(w,t)) \leq C. \tag{3.8}$$

  Since $vw \in E_j$ and $d_j(w,t) \leq d_{\ell-1}(w,t)$ by Lemma 3.7, we also have

  $$\theta + d_j(v,t) \leq \theta + \tau_{vw} + d_{\ell-1}(w,t). \tag{3.9}$$

  Thus

  $$\begin{aligned}
  \pi_v(\theta) &\geq C - \alpha d_j(v,t) - \rho(\theta + d_j(v,t)) \\
  &\geq C - \alpha d_j(v,t) - \rho(\theta+\tau_{vw}+d_{\ell-1}(w,t)) - \alpha\left(\tau_{vw} + d_{\ell-1}(w,t) - d_j(v,t)\right) \\
  &\geq \alpha d_{\ell-1}(s,t) - \alpha\tau_{vw} - \alpha d_{\ell-1}(w,t) && \text{by (3.8)} \\
  &= -\alpha\tau_{vw} - \alpha d_\ell(w,s) && \text{by Lemma 3.8} \\
  &= \pi_w(\theta+\tau_{vw}) - \alpha\tau_{vw}
  \end{aligned}$$

  where the second inequality follows from the growth bound and from (3.9).

- **Case 2:** $\pi_w(\theta + \tau_{vw}) = C - \alpha d_\ell(w,t) - \rho(\theta + \tau_{vw} + d_\ell(w,t))$.

  If $\ell \geq j$, since $vw \in E_j$ and $d_j(w,t) \leq d_\ell(w,t)$ by Lemma 3.7, we have that

  $$\theta + d_j(v,t) \leq \theta + \tau_{vw} + d_\ell(w,t). \tag{3.10}$$

  As a consequence, exploiting also the growth bound, we have

  $$\begin{aligned}
  \pi_v(\theta) &\geq \bar{\pi}_v(\theta) \\
  &= C - \alpha d_j(v,t) - \rho(\theta + d_j(v,t)) \\
  &= C - \alpha d_j(v,t) - \rho(\theta + \tau_{vw} + d_\ell(w,t)) + \rho(\theta + \tau_{vw} + d_\ell(w,t)) - \rho(\theta + d_j(v,t)) \\
  &\geq C - \alpha d_j(v,t) - \rho(\theta + \tau_{vw} + d_\ell(w,t)) - \alpha\left(\tau_{vw} + d_\ell(w,t) - d_j(v,t)\right) \\
  &= C - \rho(\theta + \tau_{vw} + d_\ell(w,t)) - \alpha\tau_{vw} - \alpha d_\ell(w,t) \\
  &= \pi_w(\theta + \tau_{vw}) - \alpha\tau_{vw}.
  \end{aligned}$$

  If $\ell < j$, by definition of $J(w, \theta + \tau_{vw})$ we have that

  $$\alpha d_\ell(s,t) + \rho(\theta + \tau_{vw} + d_\ell(w,t)) > C. \tag{3.11}$$

  Thus

  $$\begin{aligned}
  \pi_v(\theta) &\geq -\alpha d_j(v,s) \\
  &\geq -\alpha d_j(w,s) - \alpha\tau_{vw} && \text{since } vw \in E_j \\
  &\geq -\alpha d_{\ell+1}(w,s) - \alpha\tau_{vw} && \text{by Lemma 3.7} \\
  &> C - \alpha d_\ell(s,t) - \rho(\theta + \tau_{vw} + d_\ell(w,t)) - \alpha d_{\ell+1}(w,s) - \alpha\tau_{vw} && \text{by (3.11)} \\
  &= C - \alpha d_\ell(w,t) - \rho(\theta + \tau_{vw} + d_\ell(w,t)) - \alpha\tau_{vw} && \text{by Lemma 3.8} \\
  &= \pi_w(\theta + \tau_{vw}) - \alpha\tau_{vw}.
  \end{aligned}$$

- **Case 3:** $\pi_w(\theta + \tau_{vw}) = 0$.

  By Lemma 3.9, we have $\pi_v(\theta) + \alpha\tau_{vw} \geq \pi_v(\theta + \tau_{vw})$, which is nonnegative by the definition of π_v. □

This completes the proof that π satisfies all conditions of Theorem 3.4 with respect to the flow over time f produced by the algorithm, hence demonstrating the optimality of the algorithm.

3.5 Optimal tolls

Tolls $\mu : E \times \mathbb{R} \to \mathbb{R}_+$ are per-arc, time-varying and nonnegative. The value $\mu_e(\theta)$ represents the toll a user is charged upon entering the link at time θ.

We have the following theorem.

Theorem 3.11. *Let (f, π) be an optimal primal-dual solution to (3.4) (as constructed in Section 3.3 and Section 3.4) and define, for each $vw \in E$,*

$$\mu_{vw}(\theta) = (\pi_w(\theta + \tau_{vw}) - \pi_v(\theta) - \alpha\tau_{vw})^+.$$

Then f is a dynamic equilibrium under tolls μ.

Of course, to make sense of this theorem we must know what is meant by a dynamic equilibrium under tolls. Informally, it means that no user (represented as an infinitesimal flow particle) has an alternative strategy (route choice and departure time combination) of strictly smaller disutility. Making this precise in general requires defining precisely what disutility a user would incur for any given route and departure time choice, by considering the full game-theoretic Vickrey bottleneck congestion model [Vickrey, 1969, Koch and Skutella, 2011]. Tolls and departure time choice can be introduced into the definition of a dynamic equilibrium discussed in these works. However, the complexity of this is only needed because in general, a user may have to incur additional waiting time on arcs that are fully utilized, meaning that the disutility of a particular strategy depends in a complicated way on the actions of the other users. Instead we show that *even* if a user is allowed to traverse any link at any time—as if the other users were not present—there is no incentive to deviate. This is clearly a stronger property than any reasonable notion of equilibrium.

Let C be the cost horizon associated with (f, π). We will show that all users experience a disutility of exactly C with f under tolls μ, and in addition that the disutility of any other possible choice is at least C. To see this, consider any s-t-path P in G, along with departure times θ_v for each $v \in P$ valid for this path, meaning that $\theta_w = \epsilon + \theta_v + \tau_{vw}$ for every $vw \in P$, with $\epsilon \geq 0$. (So we allow the possibility of waiting at a vertex). Thus by the definition of μ and by properties (i) and (ii) of Theorem 3.4,

$$\mu_{vw}(\theta_v + \epsilon) + \alpha(\tau_{vw} + \epsilon) \geq \pi_w(\theta_w) - \pi_v(\theta_v + \epsilon) + \alpha\epsilon \geq \pi_w(\theta_w) - \pi_v(\theta_v) \ ,$$

with equality if $wv \in E^f(\theta_w)$ and $\epsilon = 0$. Then the disutility of a user using this route is

$$\begin{aligned} &\rho(\theta_t) + \sum_{e=vw \in P} [\alpha(\theta_w - \theta_v) + \mu_e(\theta_w - \tau_e)] \\ &\geq \rho(\theta_t) + \pi_t(\theta_t) - \pi_s(\theta_s) \\ &= \rho(\theta_t) + (C - \rho(\theta_t))^+ \\ &\geq C. \end{aligned}$$

The inequalities are all tight if for all $vw \in P$, $\theta_w = \theta_v + \tau_{vw}$ and $f_{vw}(\theta_v) > 0$, by the previous observations as well as property (iv). So if the aggregate choices of the users are described by f, all users pay exactly C. This completes the proof of Theorem 3.11

3.5.1 Strong vs weak enforcement

As already discussed, we cannot in general strongly enforce an optimal flow, i.e., set tolls such that every dynamic equilibrium is optimal. The following shows that the "lane tolling" approach suffices to do this.

Theorem 3.12. *Let f, π and μ be as in the previous theorem, and suppose g is any dynamic equilibrium satisfying $g_e(\theta) \leq f_e(\theta)$ for all $e \in E$, $\theta \in \mathbb{R}$. Then g is optimal.*

Proof. The cost of g cannot exceed the cost of f, and so it must be optimal. □

Essentially, being able to dynamically split and separately toll the capacity of a link allows us to easily rule out all other potential equilibria just by using tolls to artificially constrict the capacities (in addition to choosing tolls that weakly enforce the desired flow, which is still needed). Tolling in this way seems quite distant from what could be imaginable in realistic traffic scenarios. But it does raise the interesting question of whether there is a tolling scheme which can strongly enforce an optimum flow, but which is more restricted (and more plausible) than fully dynamic lane tolling. Another natural question would be to determine if an optimum flow can be strongly enforced using lane tolling only on certain specified arcs. We leave these as open questions.

3.5.2 Exogenous demand

Now let us consider the case of exogenous, and fixed, demand. Users depart from the source s at a fixed rate u_0 over a time interval $[0, T]$, and simply wish to reach the destination t as early as possible. Note that there is no longer any departure time choice; as such, users departing at different times need not experience the same disutility in an equilibrium.

We can view this within our setting as follows. Let $G' = (V', E')$ be the instance obtained by adding a node s' and an arc $s's$ of capacity u_0 and delay 0; s' becomes the new source. The total flow to send is $Q = Tu_0$. The arc $s's$ ensures that the amount of flow departing s by time θ in any flow over time cannot exceed $u_0\theta$. A flow over time need not saturate the arc $s's$ in the interval $[0, T]$; however, we can easily convert it to one that does, by simply adjusting the flow to send flow from s' to s earlier, saturating $s's$ on $[0, T]$, and then waiting at s (that is, we do not modify the flow on any other arcs). This clearly has no impact on arrival times. As such, we can view the restriction to G of any flow over time on G' as a potential solution to the exogenous demand problem.

Now let $\bar{G}$ be obtained from G' by reversing all arcs. We consider now the source to be t and the sink s', and the scheduling cost function described in (3.1). Let $\bar{f}$ be an optimal flow of value $Q = Tu_0$, and let $\bar{\mu}$ be the optimal tolls from Theorem 3.11 that induce it.

We now "reverse time". For any arc $vw \in E'$, and $\theta \in \mathbb{R}$, let $f_{vw}(\theta) = \bar{f}_{wv}(\tau_{vw} - \theta)$, and let $\mu_{vw}(\theta) = \bar{\mu}_{wv}(\tau_{vw} - \theta)$. Then f is an earliest arrival flow in G': since $\bar{f}$ minimizes the average departure time from t in $\bar{G}$, f minimizes the average arrival time at t in G'.

Lemma 3.13. *The tolls μ restricted to E induce the restriction of f to G as a dynamic equilibrium under exogenous demand.*

Proof. First, we observe that the tolls μ induce f as a dynamic equilibrium in G'. This is simply because given any s'-t-path P and valid departure times $(\theta_v)_{v\in P}$, these can be mapped to a reversed path $\bar{P}$ from t to s' in $\bar{G}$, along with corresponding reversed departure times $(\bar{\theta}_v)_{v\in\bar{P}}$ (given by, for any $vw \in P$, $\bar{\theta}_w = \tau_{vw} - \theta_v$). The disutility experienced by a user in G' choosing the strategy described by P and $(\theta_v)_{v\in P}$ is then precisely equal to the disutility experienced by a user in $\bar{G}$ choosing the strategy described by $\bar{P}$ and $(\bar{\theta}_v)_{v\in P}$.

All that remains is to go from G' to G; that is, we need to argue that the restriction of μ to V does induce the restriction of f to G. The role of $\mu_{s's}$ in G' is only to ensure equal costs between particles that traverse $s's$ at different times. This is not a requirement of an equilibrium in the exogenous setting. Some care is required however, since by restricting f to G, we are possibly introducing waiting at s.

Let $\theta' = \inf\{\theta \geq 0 : f_{s's}(\theta) < u_0\}$. Then $\mu_{s's}(\theta) = 0$ for all $\theta > \theta'$ with $f_{s's}(\theta) > 0$. To see this, consider any $\tilde{\theta} \in (\theta, \theta')$ for which $f_{s's}(\tilde{\theta}) < u_0$. Then $s's \in E^f(\tilde{\theta})$, implying by property (ii) of the dual solution π (see Theorem 3.4) that $\mu_{s's}(\tilde{\theta}) = 0$. But then if $\mu_{s's}(\theta')$ were larger than 0, it would be an improving deviation to traverse $s's$ at time $\tilde{\theta}$ and then wait at s, so this is not possible.

It follows that there is no waiting at s for users departing before time θ' in G (since f had no waiting, and $f_{s's}(\theta) = u_0$ until time θ'), whereas *all* users departing after time θ' experience the same disutility. Thus, no user has an incentive to deviate, and we have a dynamic equilibrium in G. □

3.6 General scheduling costs

We now consider general scheduling costs, satisfying only the growth bound as well as the following fairly unrestrictive condition. We will assume that for any C, $\{\theta \in \mathbb{R} : \rho(\theta) \leq C\}$ consists of a finite number of compact intervals, and this number is uniformly bounded by some value K. Insisting that this set has finite measure ensures that the total mass associated with any given choice of cost horizon is finite. The assumption that this set is always closed, or in other words, that ρ is lower semicontinuous, is a matter of convenience, and was already assumed in the strongly unimodal case. Given a scheduling function that does not satisfy this, but with a finite number of discontinuities, the property can be obtained by adjusting ρ only at points of discontinuity, and without affecting the optimal solution. Finally, the assumption that the number of intervals is bounded by some K ensures that the algorithm has a finite description, and also rules out various pathological choices of ρ.

In order to actually implement the algorithm, oracle access to ρ will not suffice. Instead, we assume that given C, we are able to obtain the sets $\rho^{-1}((-\infty, C])$ and $\rho^{-1}(\{C\})$, described as collections of intervals. Note that $\rho^{-1}(\{C\})$ consists of a union of at most $2K$ intervals, since $\rho^{-1}(\{C\}) = \rho^{-1}((-\infty, C]) \setminus \bigcup_{\epsilon>0} \rho^{-1}((-\infty, C - \epsilon])$.

There are essentially two separate complications that arise compared to the strongly unimodal case. The first complication is that the set of arrival times where the scheduling cost is bounded by some value need no longer be an interval. The second is that the total mass corresponding to a given cost horizon C need no longer depend continuously on C, meaning that once we have found the "correct" choice of cost horizon, the algorithm as stated might send too much mass.

Let us begin by dealing with the first complication alone. So suppose that in addition to the stated restrictions, $\mu\big(\rho^{-1}((-\infty, C])\big)$ is a continuous function of C, where throughout this section $\mu(A)$ denotes the Lebesgue measure of a set A. The use of bisection search (or, if ρ is given as a piecewise constant function, perhaps parametrized search) to determine the correct cost horizon will thus not be affected. We need only describe how the algorithm and analysis for finding an optimal solution for a given cost horizon should be modified.

The principle of the algorithm remains identical to its description in Section 3.3. All that changes is that for a path P_j obtained from successive shortest paths, the set of times respecting the cost horizon C is no longer an interval. Thus, we let $I_j \subseteq \mathbb{R}$ be a maximal

set such that

$$\rho(\xi + d_{j-1}(s,t)) \leq C - \alpha d_{j-1}(s,t) \quad \text{for all} \quad \xi \in I_j. \tag{3.12}$$

Given our assumptions on ρ, I_j is a finite set of compact intervals. The resulting flow f has precisely the same definition as before, namely $f_{vw}(\theta) = f_{vw}^{(J(v,\theta))}$, where the definition of $J(v,\theta)$ also remains unchanged. The proof of feasibility of this flow, and the proof of its optimality, both did not depend on any way on I_j being an interval, and the existing proofs stand as written. Note that the value of the flow is $\sum_{j=1}^{m} \mu(I_j)$; as expected, our current assumptions ensure that this is a continuous function of C.

Let us now relax the condition on continuity of $C \to \mu\Big(\rho^{-1}((-\infty, C])\Big)$. The essential idea is as follows.

- The algorithm as described will find the *maximum* mass corresponding to a given cost horizon C. We can also find the *minimum* corresponding mass, by slightly adjusting the algorithm.
- Once we have determined the correct value of C via bisection or parametric search, we must choose a solution that in a sense interpolates between the minimum and maximum mass solutions. Some care is required to ensure that we have a feasible flow. In particular, consider removing flow sent along a generalized path P_j coming from the SSP decomposition for some interval of time. If an arc e is a forward arc in P_j, then removing flow on it for a certain period of time may not be possible without also removing flow from a generalized path that uses e in the opposite direction.

We now describe the algorithm. As previously, a bisection search (or possibly parametrized search) is used to find the correct cost horizon. However, given a current guess C, we will compute a corresponding interval $[Q^{\min}(C), Q^{\max}(C)]$ of possible total masses corresponding to this. To do this, we construct, for each path P_j obtained from successive shortest paths, two sets $I_j^{\min}$ and $I_j^{\max}$ defined as follows. $I_j^{\max}$ is defined precisely as before, i.e., according to (3.12). $I_j^{\min}$ is defined instead as

$$I_j^{\min} := \mathrm{cl}\Big(I_j^{\max} \setminus \{\xi : \rho(\xi + d_{j-1}(s,t)) = C - \alpha d_{j-1}(s,t)\}\Big),$$

where $\mathrm{cl}(A)$ denotes the closure of the set $A \subseteq \mathbb{R}$. Again, our assumptions on ρ ensure that $I_j^{\min}$ is a finite collection of compact intervals. This results in two different flows over time, $f^{\min}$ and $f^{\max}$, of values

$$Q^{\min}(C) = \sum_{j=1}^{m} \mu(I_j^{\min}) \qquad \text{and} \qquad Q^{\max}(C) = \sum_{j=1}^{m} \mu(I_j^{\max})$$

respectively. (Feasibility and optimality of $f^{\min}$ has not yet been demonstrated; we will return to this point.)

Once we have found C such that $Q \in [Q^{\min}(C), Q^{\max}(C)]$, we proceed as follows. Let $\theta_0 \in \mathbb{R}$ be a value we will choose later. For each $j \in [m]$, let

$$I_j^{\theta_0} := I_j^{\min} \cup \Big(I_j^{\max} \cap (-\infty, \theta_0 - d_{j-1}(s,t)]\Big);$$

so $I_j^{\min} \subseteq I_j^{\theta_0} \subseteq I_j^{\max}$ (see Figure 3.5 for an example). Now take f^{θ_0} to be the flow over time obtained by sending flow on path P_j for times in $I_j^{\theta_0}$, for each j. Delaying concerns about feasibility, the value of this flow is $Q(\theta_0) := \sum_{j=1}^{m} \mu(I_j^{\theta_0})$. Since this is continuous and piecewise linear in θ_0, with

$$Q^{\min}(C) = \inf_{\theta} Q(\theta) \quad \leq \quad Q \quad \leq \quad \sup_{\theta} Q(\theta) = Q^{\max}(C),$$

we can easily determine the correct choice for θ_0 so that $Q(\theta_0) = Q$. The output of the algorithm is then $f := f^{\theta_0}$ for this choice of θ_0.

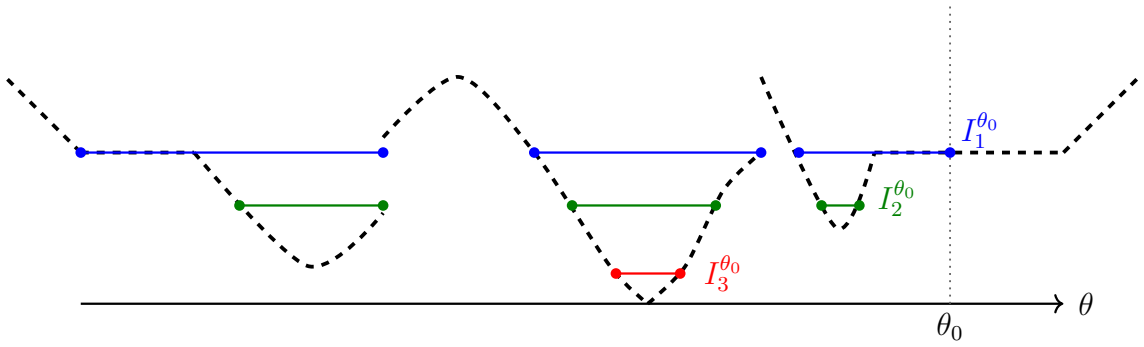


Figure 3.5: An example of a general scheduling cost function, and intervals $I_j^{\theta_0}$ corresponding to three paths of different lengths, for some particular choice of C and the indicated value of θ_0.

It remains to show the correctness of this algorithm, by demonstrating that the flow over time f constructed by this algorithm is both feasible and optimal. This could be done by suitably tweaking the arguments in Section 3.3 and Section 3.4. We will however avoid this, and instead proceed as follows. Suppose we are able to define a family of perturbed scheduling cost functions $\rho^{(\epsilon)}$ and flows $f^{(\epsilon)}$ for all $\epsilon > 0$ such that the following hold:

(i) $\rho^{(\epsilon)}$ converges uniformly to ρ as $\epsilon \to 0$. Note that this implies that the cost of an optimal solution under scheduling cost $\rho^{(\epsilon)}$ converges to the cost of an optimal solution under ρ.

(ii) $f^{(\epsilon)}$ is an optimal flow of *maximum* total mass with cost horizon C, with respect to the cost function $\rho^{(\epsilon)}$, and $f^{(\epsilon)}$ converges to f as $\epsilon \to 0$.

Together, this implies feasibility and optimality of f for ρ.

We define $\rho^{(\epsilon)}$ as follows:

$$\rho^{(\epsilon)}(\theta) = \begin{cases} \rho(\theta) & \text{if } \theta \leq \theta_0 \\ \rho(\theta) + \epsilon & \text{if } \theta > \theta_0 \\ \min\{\rho(\theta_0), \lim_{\theta' \to \theta_0^+} \rho(\theta') + \epsilon\} & \text{if } \theta = \theta_0 \end{cases}.$$

(That is, we increase ρ by ϵ to the right of θ_0, choosing the value at θ_0 so that $\rho^{(\epsilon)}$ is lower semicontinuous.) Property (i) is then immediate. Property (ii) follows from (3.12) applied to $\rho^{(\epsilon)}$: the maximal interval corresponding to path P_j in $\rho^{(\epsilon)}$ converges to $I_j^{\theta_0}$.

3.7 Conclusion

In this chapter we studied the optimization problem of finding a flow over time of value Q with minimum total cost, where the cost of each flow particle equals the journey time cost plus the scheduling cost (as discussed in Section 1.3). This objective function is very popular and frequently studied in the traffic and transportation literature but had not been studied in the optimization literature yet.

As a first contribution, we showed how to modify the successive shortest paths algorithm for earliest arrival flows in order to compute an optimal flow for this new setting. The key difference with the algorithm for earliest arrival flows lies in the new concept of computing and using a cost horizon C rather than a time horizon. However, our proof of optimality is completely different than the ones for earliest arrival flows and it is based on duality of an infinite dimensional LP.

As a second main contribution, we showed how to define, providing explicit formula, time-varying tolls that induce the social optimum as an equilibrium flow. These tolls are derived from the optimal dual solution and apply in both the endogenous and the exogenous model. This contribution is likely to be of independent interest to the transportation science community.

Since an optimal solution, both in terms of flows and corresponding inducing tolls, is based on the successive shortest paths, it possesses a clear combinatorial structure. This structure might be useful to investigate different research questions. As an example, the question of proving bounds on the the price of anarchy – i.e. the worst possible ratio between the total cost of an equilibrium and the one of a social optimum.

We will use these results in Chapter 5, where we will compare the equilibria with optimal tolls, i.e. social optima, with the untolled ones.

Chapter 4

Dynamic equilibria with endogenous departure time choice

This chapter contains unpublished results obtained thanks to the collaboration of Neil Olver.

4.1 Introduction

In this chapter we define the model of *dynamic equilibria with endogenous departure time choice* and we study their existence and uniqueness. Here users are able to choose both their route and their departure time. They are then concerned not only with their journey time, but also their *arrival time* at the destination and, in addition to the journey time costs, incur also scheduling costs for arriving either early or late (see Section 1.3). A dynamic equilibrium (or simply equilibrium) occurs then when no user has a unilateral deviation that strictly decreases their disutility. This implies that the inflow rate into the network it is not part of the input, as it is in the exogenous demand model, but depends on the equilibrium behavior.

We represent traffic dynamics utilizing the Vickrey bottleneck congestion model [Vickrey, 1969] (see Section 1.1) and traffic flow utilizing the flow over time model with waiting on the arcs defined in Section 2.4. Here each user is represented by a divisible infinitesimally small particle that possesses no mass. We focus on an homogeneous setting where users share origin, destination and scheduling cost function and therefore a setting where all the users share the same strategy set (i.e., the same collection of routes and departure times). This implies that all the agents incur the same cost.

In fact, instead of specifying the total mass of users Q, we consider a cost horizon C that represents the cost that the agents are willing to pay. That is, an agent will commute if their total cost for doing so is strictly below C, and is indifferent to commuting or not if this cost is exactly C, but will not commute at any larger cost.

While in the previous chapter we considered any scheduling cost function, in this chapter we focus only on convex scheduling cost functions ρ for which $\theta \to \rho(\theta) + \alpha\theta$ is increasing. The latter condition, that we call the *growth bound* on ρ and that is similar to the condition of the same name of Chapter 3[1], is a necessary condition for the existence of

[1]With the difference that the latter was nondecreasing instead of increasing.

the equilibrium with arbitrary cost horizon C. We postpone this discussion to Section 4.2.

This setting (with the more restricted scheduling cost function defined in Equation (1.1)) is very common in transportation literature (see [Small, 2015, Li et al., 2020] for a survey) but most of the attention received has been on simple networks, like single-link, parallel-links networks or networks with very few arcs [Arnott et al., 1993, Kuwahara, 1990, Zhang et al., 2008]; in this chapter we consider arbitrary network topologies, which have received much less attention. The perspective we take here follows the treatments in the optimization literature (e.g., [Cominetti et al., 2015, Koch and Skutella, 2011]), where users do not choose their departure time but are released into the network at a given inflow rate.

Dynamic equilibria with endogenous departure time choice are used in the rest of the thesis except for Chapter 7.

Our results and outline of the chapter. We formally define the model in Section 4.2. We will refer to this section also in Chapter 7, where we will study dynamic equilibria with exogenous departure time choice.

In Section 4.3 we show existence and uniqueness of the equilibrium in a *sensitive* demand model, where each user have some fixed disutility that they are willing to suffer, and otherwise will not travel. Then, in Section 4.4, under two very plausible conjectures about continuity and monotonicity of equilibria, we extent the existence and uniqueness results to the insensitive model.

Finally, in Section 4.5 we show how dynamic equilibria with endogenous departure time choice generalize the ones where users do not choose the departure time.

4.2 Model and preliminaries

As mention earlier, in this chapter we follow the methodology used in the optimization literature (e.g., [Cominetti et al., 2015, Koch and Skutella, 2011]) and many proofs rely on the same concepts and ideas. We refer to Chapter 2 for the network structure, notation and flow over time definition.

Link dynamics. In this model we consider flow over time with waiting on arcs and not on vertices. This means that strict flow conservation is satisfied and that, if the inflow rate into an arc exceeds the capacity, a *queue* forms at the entrance. We use $z_e(\theta)$ to denote the total mass of the queue on arc e at time θ, which equals the total amount of flow that has entered e by time θ minus the total amount of flow that has exit e by time $\theta + \tau_e$, i.e.:

$$z_e(\theta) = F_e^{in}(\theta) - F_e^{out}(\theta + \tau_e) \ .$$

Queues are *vertical* and *spaceless*, in the sense that they possess no physical dimension and can become arbitrarily large without interfering with traffic not using the arc in question (there is no *spillback*). So one can equivalently think of the queue as being formed at the exit of the arc. Additionally, users arrive at the end of the queue instantaneously and independently on the size of it.

Once inside the network, users want to arrive at their destination as soon as possible, a feature that results from the growth bound of the scheduling cost function. This implies

that users leave a queue according to the First-In-First-Out (FIFO) principle and that queues evacuate at the maximum possible rate, i.e. the capacity ν_e. As a result, a particle that enters the arc at time θ experiences a queueing delay of

$$q_e(\theta) := z_e(\theta)/\nu_e \ . \tag{4.1}$$

The time at which the particle exits the arc equals thus the sum of the entrance time into the arc, the queueing delay and the free transit time of the arc:

$$T_e(\theta) := \theta + q_e(\theta) + \tau_e. \tag{4.2}$$

The evolution of the queue is then described via the link dynamics

$$\frac{d}{d\theta} z_e(\theta) = \begin{cases} f_e^{in}(\theta) - \nu_e & \text{if } z_e(\theta) > 0 \\ \max\{f_e^{in}(\theta) - \nu_e, 0\} & \text{if } z_e(\theta) = 0. \end{cases} \tag{4.3}$$

The outflow rate at time $\theta + \tau_e$ is also clear: if there is no queue, it is equal to the inflow rate, and if there is a queue, then it is equal to the capacity.

$$f_e^{out}(\theta + \tau_e) = \begin{cases} f_e^{in}(\theta) & \text{if } z_e(\theta) = 0 \\ \nu_e & \text{if } z_e(\theta) > 0. \end{cases} \tag{4.4}$$

Given a collection of inflow functions $(f_e^{in})_{e\in E}$, the corresponding outflow functions f_e^{out} and queue functions z_e can be deduced from (4.3) and (4.4). Therefore we specify a flow over time by just using its inflow functions $(f_e^{in})_{e\in E}$.

Earliest arrival functions and dynamic shortest path network. Consider some flow over time $(f_e^{in})_{e\in E}$, we define the *earliest arrival functions* ℓ_v, for $v \in V$, such that $\ell_v(\theta)$ is equal to the earliest time a particle can arrive at v, given that it leaves s at time θ. Let us be more precise and let's define $\ell_v(\theta)$ using the dynamic Bellman's equations

$$\begin{cases} \ell_s(\theta) = \theta & \forall \theta \in \mathbb{R} \\ \ell_w(\theta) = \min\limits_{vw\in E} T_{vw}(\ell_v(\theta)) & \forall w \in V \setminus \{s\}, \theta \in \mathbb{R} \ . \end{cases} \tag{4.5}$$

Let the *entrance time* of a particle be the time this particle departs from s. We say that an arc $e = vw$ is *active* at entrance time θ if $\ell_w(\theta) = T_e(\ell_v(\theta))$. This means that it is possible for a particle leaving s at time θ to arrive at w as early as possible, namely at time $\ell_w(\theta)$, by a path that includes the arc e. Moreover, if the active arc $e = vw$ hosts a queue at time $\ell_v(\theta)$, we call it *resetting* at entrance time θ. Let $E'_\theta, E^*_\theta \subseteq E'_\theta$ be, respectively, the set of all the active arcs and the set of all the resetting arcs at entrance time θ, formally:

$$E'_\theta := \{e = vw \in E : \ell_w(\theta) = T_e(\ell_v(\theta))\}$$

and

$$E^*_\theta := \{e = vw \in E'_\theta : \ell_w(\theta) > \ell_v(\theta) + \tau_e\} \ .$$

The network $G_\theta = (V, E'_\theta, E^*_\theta)$ is called the *shortest path network at time* θ. Note that G_θ is acyclic since the complete network has no directed cycle with zero transit time (see Section 2.2) and because of Equation (4.5).

User costs and choices. Each individual user is considered to control a negligible fraction of the total flow, so that there are an infinite number of infinitesimally small, divisible users. We think of each user as being able to freely choose both their departure time and their route through the network. The joint choices of all these users should then induce a corresponding flow over time. We restrict our attention to a setting with completely homogeneous users, with origin s, destination t and desired arrival time T^*.

As in in Chapter 3, we consider a scheduling cost setting: the total disutility of a user is the sum of their scheduling cost and their journey time, scaled by some factor $\alpha > 0$ representing his value of time (see Section 1.3). We use $\rho(\theta)$ to denote the scheduling cost associated with arriving at time θ. The cost of a particle with entrance time θ is thus

$$\alpha \cdot \Big(\ell_t(\theta) - \theta\Big) + \rho\Big(\ell_t(\theta)\Big) \ . \tag{4.6}$$

Dynamic equilibria and induced cumulative flow. A *dynamic equilibrium* (or simply *equilibrium*) is a joint choice amongst all the users in the system of routes and departure times, with the property that no user has a unilateral deviation that strictly decreases their disutility. This results in a flow over time which we call *equilibrium flow*. Since all users have the same strategy set (i.e., the same collection of routes and departure times) and are homogeneous, all users should experience the same disutility in a deterministic equilibrium. Then the following condition is a requirement for a flow to be an equilibrium:

Condition 4.1. *For all $e = vw \in E$ and almost every θ for which $f_e^{in}(\ell_v(\theta)) > 0$, e is active at entrance time θ.*

Condition 4.1 says that the equilibrium flow uses only paths in the current shortest path network and ensures that all users are satisfied with their choice of route, given their choice of departure time. This can be expressed through the following lemma:

Lemma 4.2 ([Cominetti et al., 2015, Theorem 1])**.** *If a flow over time $(f_e^{in})_{e\in E}$ is a dynamic equilibrium then $F_e^{in}(\ell_v(\theta)) = F_e^{out}(\ell_w(\theta))$ for each arc $e = vw \in E$ and $\theta \in \mathbb{R}$.*

It follows that the functions $x_e(\theta) := F_e^{in}(\ell_v(\theta))$, called *cumulative flow induced by a dynamic equilibrium*, are static flows with

$$\sum_{e\in\delta^{out}(v)} x_e(\theta) - \sum_{e\in\delta^{in}(v)} x_e(\theta) = \begin{cases} 0 & \text{if } v \neq s,t \\ \int_{-\infty}^{\theta} \nu_0(\xi)d\xi & \text{if } v = s, \end{cases} \tag{4.7}$$

where $\nu_0(\theta)$ is the (non-negative) rate of users departing from s at time θ (note that a user might "depart" s at some moment but then immediately be forced to wait on an arc).

We must add to Condition 4.1 a condition that ensures all users are satisfied with their choice of departure time as well. Thus we have

Condition 4.3. *For any θ and θ', where in addition $\nu_0(\theta) > 0$, we have*

$$\alpha\Big(\ell_t(\theta) - \theta\Big) + \rho\Big(\ell_t(\theta)\Big) \leq \alpha\Big(\ell_t(\theta') - \theta'\Big) + \rho\Big(\ell_t(\theta')\Big).$$

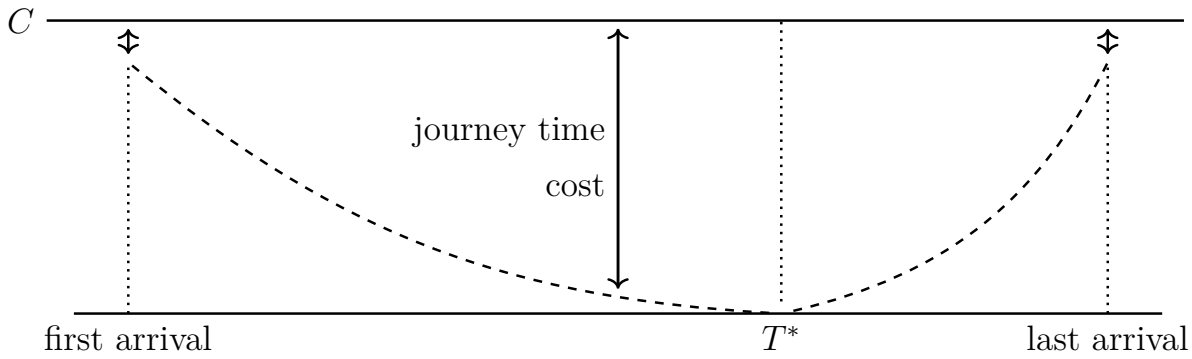


Figure 4.1: The horizontal axis represents the arrival time. The dashed line indicates the scheduling cost function. The top black line indicates the cost of an agent. The vertical distance between the top black line and the dashed line indicates the journey time cost for an agent arriving at time indicated by the horizontal axis.

Let T^* denote the time when the scheduling cost is minimum, i.e. $T^* = \operatorname{argmin}_\theta \rho(\theta)$. Condition 4.3 implies that all the particles of the equilibrium flow have the same cost. This means that the particles that arrive at the sink first incur a small journey time cost but a big scheduling cost. On the other hand, the particles that arrive at the sink closely to time T^* incur a small scheduling cost but a big journey time cost (see Figure 4.1).

Additionally, let the *starting-time* and the *ending-time*, denoted as θ_f and θ_l, be the departure time of, respectively, the first and last particle that arrives at t. Note that Condition 4.1 implies that the particle that arrives first (last) at t is also the particle that departs first (last) from s. Since ρ is continuous and convex, θ_f and θ_l are uniquely defined such that $\theta_f \leq \theta_l$, $\rho(\theta_f + \tau_{\text{SP}}) = C - \alpha \cdot \tau_{\text{SP}}$ and $\rho(\theta_l + \tau_{\text{SP}}) = C - \alpha \cdot \tau_{\text{SP}}$, where C denotes the cost horizon of the equilibrium flow, i.e. the total cost of any user.

A consequence of Condition 4.3 is the following lemma:

Lemma 4.4. *For the earliest arrival functions $(\ell_v)_{v \in V}$ of an equilibrium we have that $\ell_t'(\theta) = \frac{\alpha}{\alpha + \rho'(\ell_t(\theta))}$ for $\theta \in (\theta_f, \theta_l)$.*

Proof. Let's consider two particles that depart at time θ and $\theta + \varepsilon$, where $\varepsilon > 0$ and $\theta, \theta + \varepsilon \in (\theta_f, \theta_l)$. Since the two particles incur the same cost, it follows that:

$$\alpha\Big(\ell_t(\theta) - \theta\Big) + \rho\Big(\ell_t(\theta)\Big) = \alpha\Big(\ell_t(\theta+\varepsilon) - (\theta+\varepsilon)\Big) + \rho\Big(\ell_t(\theta+\varepsilon)\Big)$$
$$\alpha\Big(\theta + \varepsilon - \theta\Big) = \alpha\Big(\ell_t(\theta+\varepsilon) - \ell_t(\theta)\Big) + \rho\Big(\ell_t(\theta+\varepsilon) - \ell_t(\theta)\Big)$$

taking the limit as ε goes to zero, the statement follows. □

Growth bound of the scheduling cost function. As already discussed, we assume that ρ satisfies the growth bound, i.e that $\theta \to \rho(\theta) + \alpha\theta$ is increasing.

This condition, as Lemma 4.4 suggests, is necessary for the existence of the equilibrium with arbitrary cost horizon C. Consider a scheduling cost function ρ where $\rho'(\theta) \leq -\alpha$ for some maximal interval $I = [\theta_1, \theta_2)$. This implies that the particles arriving at destination

at time $\theta \in I$ could "wait" at the sink until time θ_2 without increasing their scheduling cost and therefore that all the particles that arrive at t within the time interval I have the same scheduling cost. But then, since in an equilibrium everyone has the same total cost (by Condition 4.3), all these particles have also the same journey time cost. This happens only if they all depart and arrive at the same time. Additionally, note that these particles encounter no queue delays because, otherwise, they should arrive at destination immediately after the agents that arrive at destination before time θ_1.

As a consequence, there cannot be a cost horizon larger than $\rho(\theta_2) + \alpha\tau_{\mathrm{SP}}$, where τ_{SP} is the travel time of the shortest s-t-path.

Derivatives of a dynamic equilibrium. For almost every $\theta \in (\theta_f, \theta_l)$, let $x' := \frac{dx}{d\theta}$ and $\ell' := \frac{d\ell}{d\theta}$ be the derivatives of the cumulative flow $(x_e)_{e\in E}$ and the earliest arrival functions $(\ell_v)_{v\in V}$. By differentiating Equation (4.7) we see that $x'(\theta)$ is a static flow. Moreover, by the Bellman equations (4.5), by Lemma 4.4 and by the equilibrium conditions (4.1), (4.2), (4.3) and (4.4) we have that x', ℓ' fulfill the following definition.

Definition 4.5 (Endogenous Thin Flow with Resetting (ETF))**.** *We say that the pair (x', ℓ') is an Endogenous Thin Flow on $G_\theta = (V, E'_\theta, E^*_\theta)$ if:*

$$
\begin{aligned}
&x' \text{ is a static flow on } (V, E'_\theta) && \\
&\ell'_s = 1 && \\
&\ell'_t = \frac{\alpha}{\alpha + \rho'(\ell_t(\theta))} && \\
&\ell'_w = \min_{vw \in E'_\theta} \eta(\ell'_v, x'_{vw}) && \forall v \in V \setminus \{s\} \\
&\ell'_w = \eta(\ell'_v, x'_{vw}) && \forall vw \in E'_\theta \text{ with } x'_{vw} > 0
\end{aligned}
$$

where

$$
\eta(\ell'_v, x'_{vw}) = \begin{cases} \frac{x'_{vw}}{\nu_{vw}} & vw \in E^*_\theta \\ \max\{\ell'_v, \frac{x'_{vw}}{\nu_{vw}}\} & vw \notin E^*_\theta. \end{cases}
$$

Note that the definition of endogenous thin flow with resetting matches the one of the *Normalized Thin Flows* introduced in [Cominetti et al., 2015], with the only difference that the latter fixes the inflow into s while the former fixes the value of ℓ'_t. We reproduce it here for convenience.

Definition 4.6 (Normalized Thin Flow with Resetting (NTF)[Cominetti et al., 2015])**.** *We say that the pair (x', ℓ') is a Normalized Thin Flow on $G_\theta = (V, E'_\theta, E^*_\theta)$ if:*

$$
\begin{aligned}
&x' \text{ is a static flow on } (V, E'_\theta) \text{ with value } \nu_0 && \\
&\ell'_s = 1 && \\
&\ell'_w = \min_{vw \in E'_\theta} \eta(\ell'_v, x'_{vw}) && \forall v \in V \setminus \{s\} \\
&\ell'_w = \eta(\ell'_v, x'_{vw}) && \forall vw \in E'_\theta \text{ with } x'_{vw} > 0
\end{aligned}
$$

where

$$\eta(\ell'_v, x'_{vw}) = \begin{cases} \frac{x'_{vw}}{\nu_{vw}} & vw \in E^*_\theta \\ \max\{\ell'_v, \frac{x'_{vw}}{\nu_{vw}}\} & vw \notin E^*_\theta. \end{cases}$$

4.3 Existence and uniqueness in the sensitive demand model

In this section we prove the existence and uniqueness of equilibria with endogenous departure time choice in a *sensitive* demand model, where users travel only if and only if their disutility is below a fixed cost horizon C.

The next theorem shows the existence of the endogenous thin flow with resetting.

Theorem 4.7. *Let $G_\theta = (V, E'_\theta, E^*_\theta)$ be the current shortest path network. Unless $\ell'_t(\theta) < 1$ and $E^*_\theta = \varnothing$, there exists an ETF on G_θ. Moreover, all the ETFs on G_θ have the same labels $(\ell'_v)_{v\in V}$.*

In order to prove Theorem 4.7 we need the following Lemma.

Lemma 4.8. *Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be a function defined such that $\psi(\nu_0)$ is the value of $\ell'_t(\theta)$ of the NTFs on G_θ with inflow ν_0. If $E^*_\theta = \varnothing$ then, for any $\sigma \in [1, \infty)$, there exists a value of ν_0 such that $\psi(\nu_0) = \sigma$. If $E^*_\theta \neq \varnothing$ then, for any $\sigma \in (0, \infty)$, there exists a value of ν_0 such that $\psi(\nu_0) = \sigma$.*

Proof. For the first part of the statement consider the min-cut capacity ς of the shortest path network. From the algorithm to compute a thin flow with resetting for the special case of $E^*_\theta = \varnothing$ ([Koch and Skutella, 2011]), we know that, if $\nu_0 \leq \varsigma$ then $\ell'_t = 1$ and that if $\nu_0 = \sigma \cdot \varsigma$, for any $\sigma \geq 1$, then $\ell'_t = \sigma$.

For the second part of the statement we know that, given a constant inflow and a shortest path network G_θ, there exists a NTF on G_θ ([Cominetti et al., 2015, Theorem 3]). Our goal is hence to show that there exists an inflow ν_0 such that $\ell'_t = \sigma$, for any $\sigma > 0$.

From [Cominetti et al., 2015, Kaiser, 2020] we know that ψ is a continuous function and from [Kaiser, 2020, Theorem 14], or by slightly modifying the proof of [Cominetti et al., 2015, Theorem 4], we know that it is monotonically nondecreasing. Thus we just need to show that ψ has a minimum that is not larger than the desired value of ℓ'_t and a maximum that is not smaller than the desired value of ℓ'_t.

Let ν^* be the smallest capacity of an arc in the shortest path network, i.e. $\nu^* = \min\{\nu_e : e \in E'_\theta\}$. Since G_θ is acyclic, it follows that $x'_e \leq \nu_0$ for any arc e and that $\ell'_w \leq \frac{\nu_0}{\nu_{vw}} \leq \frac{\nu_0}{\nu^*}$ for any resetting arc vw. By the flow conservation we know that from the head of a resetting arc to t there exists a path carrying flow. This implies, from the thin flow constraints, that $\ell'_t \leq \frac{\nu_0}{\nu^*}$. As a consequence, by choosing an inflow rate ν_0 smaller than $\sigma \cdot \nu^*$ we obtain that $\ell'_t \leq \sigma$.

Now let υ indicate the sum of the capacities of the in-going arcs of t in the shortest path network, i.e. $\upsilon = \sum_{e=vt\in E'_\theta} \nu_e$. Since all the flow reaches t, there exists an arc e among the in-going arcs of t with $\frac{x'_e}{\nu_e} \geq \frac{\nu_0}{\upsilon}$. But then, by the thin flow constraints, we have that $\ell'_t \geq \frac{\nu_0}{\upsilon}$. By choosing an inflow rate ν_0 greater than $\sigma \cdot \upsilon$ we get $\ell'_t \geq \sigma$, concluding the proof. □

We are now ready to prove Theorem 4.7.

Proof of Theorem 4.7. As mention earlier, the endogenous thin flows with resetting are similar to normalized thin flows with resetting (NTFs) with the difference that the former fixes the value of ℓ'_t while the latter fixes the value of the inflow ν_0. By [Cominetti et al., 2015, Theorem 3], we know that, given a constant inflow and a shortest path network G_θ, there exists a NTF on G_θ and that all the NTFs on G_θ have the same labels $(\ell'_v)_{v\in V}$ [Cominetti et al., 2015, Theorem 4]. Therefore, in order to prove our statement, we just need to show that there is a value of ν_0 such that the resulting NTF has the desired value for ℓ'_t. This follows by Lemma 4.8. □

We remark that there does not exist an ETF (x', ℓ') with $\ell'_t < 1$ and $E^*_\theta = \varnothing$ since, in this case, we have that $\ell'_t \geq \ell'_s = 1$.

The next theorem shows existence and uniqueness for any cost horizon C. As in the case of the exogenous demand model, our uniqueness result concerns only right-continuous equilibria, i.e. equilibria whose earliest arrival functions $(\ell_v)_{v\in V}$ are right-continuous.

Theorem 4.9. *For any cost horizon C, there exists an equilibrium. Moreover, the earliest arrival functions $(\ell_v)_{v\in V}$ are the same for all equilibria with cost horizon C and that are right-continuous.*

Proof. The first part of the statement follows by Theorem 4.7 and by the *equilibrium extension procedure* of Koch and Skutella [2011], Cominetti et al. [2015]. This procedure consists in integrating the thin flows in phases during which the thin flows do not change. More precisely, these phases are maximal time intervals in which the inflow into the source is constant, no arc is added to the shortest path network and no queue totally depletes. One step of the procedure is called κ-extension[2] and formally consists in choosing the largest value of κ such that

$$\ell_w(\theta) - \ell_v(\theta) - \tau_{vw} + \kappa \cdot (\ell'_w - \ell'_v) \geq 0 \qquad \forall vw \in E^*_\theta \tag{4.8}$$

$$\ell_w(\theta) - \ell_v(\theta) - \tau_{vw} + \kappa \cdot (\ell'_w - \ell'_v) \leq 0 \qquad \forall vw \in E \setminus E'_\theta \tag{4.9}$$

where (x', ℓ') is an ETF on the shortest path network $G_\theta = (V, E'_\theta, E^*_\theta)$. Equation (4.8) ensures that no queue length becomes negative (it holds with equality when the queue completely depletes). Equation (4.9) ensures that the arcs not present in the shortest path network are unattractive (it holds with equality when the arc vw is added to the shortest path network). Once found the value of κ, the shortest path network for time $\theta + \kappa$ is computed.

To construct an endogenous equilibrium with endogenous departure time choice we can proceed similarly to the case of exogenous demand: we apply the κ-extension procedure from the starting-time θ_f (that depends on C) until the ending-time θ_l, which corresponds to the time departure of the last particle that arrives at the sink and that experiences no waiting.

Each extension is referred to as a *phase in the evolution of the equilibrium*. Note that we do not have a lower bound for the length of a phase (see [Cominetti et al., 2017]).

[2] Originally called α-extension and here renamed since the variable α indicates the value of time.

Therefore, if the number of phases is not finite, we define the equilibrium by using the point-wise limit of the labels $(\ell'_v)_{v\in V}$, which exists since the earliest arrival functions $(\ell_v)_{v\in V}$ are nondecreasing and have bounded derivatives.

The equilibrium thus constructed is right-continuous and the inflow in each phase is constant. Consequently, the uniqueness result follows from [Cominetti et al., 2015, Theorem 6]: *Suppose that the inflow into the network is piecewise constant. Then, the earliest arrival functions $(\ell_v)_{v\in V}$ are the same for all dynamic equilibria which are right-continuous.* □

4.4 Existence and uniqueness in the insensitive demand model

In this section we discuss the relation between the cost horizon C and the total mass Q of users in an equilibrium.

Consider for a moment the case where the scheduling cost function is the most common one; where users arriving at a time $\theta \leq T^*$ experience a cost of $\beta(T^*-\theta)$, and users arriving at a time $\theta > T^*$ experience a cost of $\gamma(\theta - T^*)$ (see Equation (1.1) and Figure 1.2). As a consequence of Theorem 4.9, and by the fact that the value of $\ell'_i(\theta)$ does not change for $\theta < \theta^*$, we have that the evolution of a right-continuous equilibrium with starting-time θ_f, in the time interval $[\theta_f, \theta_f + \vartheta]$, with $\theta_f + \vartheta \leq \theta^*$, matches the evolution of a right-continuous equilibrium with starting-time $\bar{\theta}_f \leq \theta_f$ in the time interval $[\bar{\theta}_f, \bar{\theta} + \vartheta]$. This implies that the amount of flow arriving at the sink before T^* continuously and monotonically increases with C, or equivalently, increases when the starting-time of the equilibrium decreases.

Unfortunately, we could not prove this monotonicity property to be true also for the amount of flow arriving at the sink after time T^* and, more generally, we could not prove it for more general convex scheduling cost functions. Nonetheless we believe it to be true. More precisely, let the function $\Upsilon : \mathbb{R}_+ \to \mathbb{R}_+$ be defined such that $\Upsilon(C)$ equals the amount of flow of a right-continuous equilibrium with cost horizon C. We believe the following conjectures to hold.

Conjecture 4.10. *Υ is a continuous function.*

Conjecture 4.11. *Υ is a monotonically nonincreasing function.*

Notice that there is then a one-to-one map between Q and the cost horizon C and, consequently, if Conjecture 4.10 holds, then we have the existence of an equilibrium of mass Q, for any $Q > 0$; additionally, if Conjecture 4.11 holds, then we have that the earliest arrival functions $(\ell_v)_{v\in V}$ are the same for all the equilibria that have mass Q and that are right-continuous.

4.5 Reduction from equilibria with endogenous departure time choice to equilibria with exogenous ones

In this section we show how the equilibria with endogenous departure time choice generalizes the ones where users depart from the source s at a given constant positive rate over a time interval $[0, T)$, and simply wish to reach the destination t as early as possible.

Given an equilibrium $(\bar{x}, \bar{\ell})$ of an instance in the exogenous model with constant inflow rate $\bar{\nu}_0$ and graph $G = (V, E)$, we will construct an instance for the endogenous model that induces an equilibrium (x, ℓ) matching the dynamics and inflow rate of $(\bar{x}, \bar{\ell})$. We achieve this by modifying G and by selecting an appropriate scheduling cost function.

Let $G' = (V', E')$ be the graph obtained from G by adding a new vertex t' and a new arc tt' of capacity ν^* and free transit time 0, where $0 < \nu^* < \min\{\nu_e, \bar{\nu}_0\}$ for any $e \in E$; t' becomes the new sink. By choosing such capacity for the new arc tt' we make sure that this arc is always resetting and thus that the label $\ell'_{t'}(\theta)$ is always equal to $\nu_0(\theta)/\nu^*$: consider an arbitrary time θ that admits an ETF, since G_θ is acyclic, we have that $x'_e(\theta) \leq \nu_0(\theta)$ for any arc e and, therefore, that $\ell'_w(\theta) \leq \max\{1, \nu_0(\theta)/\nu_e\} < \nu_0(\theta)/\nu^*$ for any arc $e = vw \in E'_\theta$. This implies that $\ell'_t(\theta) < \nu_0(\theta)/\nu^*$ and that $\ell'_{t'}(\theta) = \nu_0(\theta)/\nu^*$.

Now let $\alpha = \bar{\nu}_0/\nu^*$ be the journey time cost, τ_0 be the length of the shortest path in G and consider the following scheduling cost function

$$\rho(\theta) = \begin{cases} (\alpha - 1)(T^* - \theta) & \text{if } \theta < T^* \\ \gamma(\theta - T^*) & \text{if } \theta \geq T^* \end{cases} \tag{4.10}$$

where $T^* = \tau_0 + (\alpha \cdot T)$ and γ is an arbitrarily chosen positive value. Note that this scheduling cost function represents the standard α-β-γ scheduling cost function given in Equation (1.1) with $\beta = \alpha - 1$.

Consider now the cost horizon $C = \tau_0 + (\alpha - 1)T^*$. This cost horizon and the scheduling cost function ensure that the departure time of the first agent occurs at time 0, that $\ell'_{t'}(\theta) = \bar{\nu}_0/\nu^*$ for all points $\theta \in [0, T)$ and that the arrival time of the particle that departs at time T is T^*. We will only consider the dynamics of the particles that depart in the time interval $[0, T)$.

It follows that, in order to get $\ell'_{t'}(\theta) = \bar{\nu}_0/\nu^*$, the flow on tt' has to be equal to $\bar{\nu}_0$ and hence we have that $\nu_0(\theta) = \bar{\nu}_0$. But then there is no difference between the endogenous thin flows and the normalized thin flows for any time $\theta \in [0, T)$ in G (Definitions 4.5 and 4.6). Thus the exogenous and endogenous equilibria are essentially the same and we can ignore the dynamics on tt'.

4.6 Conclusion

In this chapter we introduced a model to represent traffic equilibria where users are concerned with both their route and their arrival time at destination. For the latter, as in Chapter 3, we considered a scheduling cost setting, with the difference that here we restricted our attention to convex functions. An equilibrium is a joint choice amongst all

the users in the system of routes and departure times, with the property that no user has a unilateral deviation that strictly decreases their cost. We also provided an algorithm to compute an equilibrium.

Additionally we showed that, under two highly plausible conjectures, the equilibrium always exists and it is unique.

This chapter represents the core of the thesis, hence the similarity of its title with the thesis' one. In fact, it is connected to Chapter 3 through the use of scheduling cost functions; the model and algorithm here introduced will be used in Chapters 5 and 6 and, finally, as we showed in Section 4.5, the model generalizes the one that will be investigated in Chapter 7.

Chapter 5

Emergent hypercongestion in Vickrey bottleneck networks

The results in this chapter were published in [Frascaria, Olver, and Verhoef, 2020].

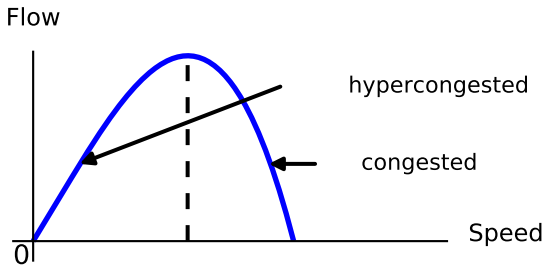
5.1 Introduction

Empirical studies have shown that traffic throughput can be lower during peak hours of high demand than during less congested hours. This phenomenon is called *hypercongestion*. Hypercongestion has been studied for single links, as well as for networks as a whole. For single links, it is understood through the *fundamental diagrams of traffic flow* (Figures 5.1a to 5.1c) that relate traffic density, traffic speed and traffic flow, and these mechanisms are of particular importance in modelling highway traffic flow. In highway traffic, it is a well-understood and intuitive fact that speed is negatively related to density. Since flow is the product of speed and density, maximum flow is achieved at some intermediate value of density. Increasing density beyond that point (or equivalently, decreasing speed) yields to decreased throughput; this is hypercongestion.

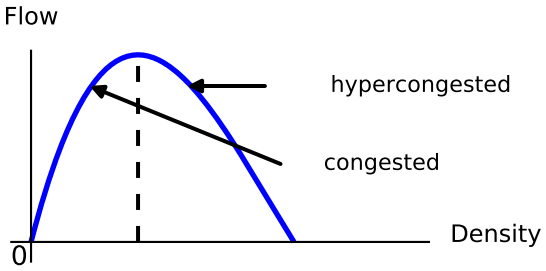
For networks as a whole, representing, for example, the downtown core of a city, "macroscopic" versions of the fundamental traffic diagram have been empirically observed (e.g., [Geroliminis and Daganzo, 2008, Daganzo et al., 2011]). Figure 5.1d reproduces Figure 9 from [Daganzo et al., 2011]: here the vertical axis represents the number of vehicles passing, per lane and per unit of time, the points of measurement and the horizontal axis represents the density of the network, which is proportional to the number of vehicles present in the network.

The effect has been discussed in the context of *bathtub models* [Arnott, 2013, Fosgerau, 2015, Arnott et al., 2016], and indeed is a key motivation for these models. Here, speed and density are *uniform* across the space, even though agents have different routes with different lengths. With this strong spatial homogeneity assumption, the network structure of the city and its road structure is abstracted away and network interactions[1] are completely absent. A negative relationship (e.g., Greenshields' linear relation [Greenshields, 1935]) is then prescribed between the (spatially averaged) density and speed. This leads

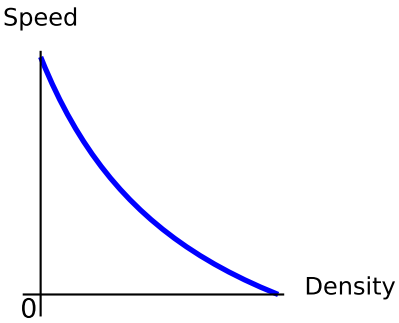
[1] Temporary and spatially separate events that affect each other.



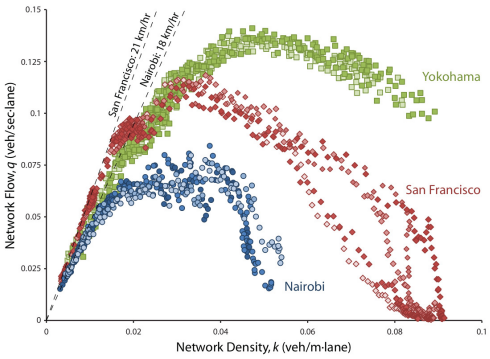
(a) Speed-flow diagram.



(b) Density-flow diagram.



(c) Speed-density (or fundamental) diagram.



(d) Macroscopic fundamental diagram from [Daganzo et al., 2011, Fig.9].

Figure 5.1: Fundamental diagrams of traffic flow.

to hypercongestion when density exceeds some critical value, for the same reason as for a single link.

While these models do an excellent job of matching empirically observed behavior such as the macroscopic fundamental diagram of Figure 5.1d, they do not provide an explanation for the source of this negative relationship, and hence the source of hypercongestion. In particular, they typically lack explicit modelling of entry into and exit out of the bathtub, while even for single-link models it has been shown that hypercongestion can only build up if exit capacity is restricted (i.e., there is a downstream bottleneck), while entry capacity should not be below the downstream exit capacity (e.g., [Verhoef, 2001, 2003]). To the extent that the bathtub is modeled as a simple congestible facility, one might expect the same type of necessary conditions for hypercongestion to occur in these models, which would justify an explicit consideration of entry and exit mechanisms. In any case, it seems worthwhile to search for a deeper explanation of the precise causes of hypercongestion in urban traffic, as this may produce new insights on optimal transport (pricing) policies.

The most obvious explanation for this relationship between averaged density and speed would be that the relationship holds at the level of individual links. In other words: if hypercongestion occurs at the level of individual links, we would expect it to occur in the macroscopic level as well. Hypercongestion at the link level can be considered as expected on highways entering the city, given that downstream capacity is limited. But *within* the city, is is less evident that one should expect substantial per-link hypercongestion.

This is especially relevant given that the Vickrey bottleneck model [Vickrey, 1969] is generally considered to be a good "workhorse" dynamic economic model for the most heavily congested links in the context of urban traffic. However, in its purest form, this model does not exhibit the flow drops that are characteristic for hypercongestion at the link level.

An alternative source for hypercongestion is through spillback (coined "triggerneck congestion" by Vickrey [1969]). If traffic congestion in one part of the network can block upstream intersections, it is relatively easy to construct examples where hypercongestion occurs.[2] In highly congested situations, the potential for "gridlock" caused by spillback effects is a very plausible mechanism for hypercongestion in cities. But what about lower levels of congestion where the impact of spillback is nonexistent or small? What about the role of dynamic route choices and network interactions before triggerneck congestion sets in and, therewith, the role of network design? Can that in itself be a source of hypercongestion at the level of the network? In light of these considerations, we ask the following natural question.

Question 5.1. Can hypercongestion occur in settings where (1) no hypercongestion occurs at the link level, in that the flow-speed relationship does not exhibit a backward-bending curve; and (2) there are no spillback effects?

Our assumptions in asserting whether dynamic network behavior can produce hypercongestion through route interaction are therefore conservative: if, in our model, hypercongestion is observed at the network level, it can be fully attributed to the network interaction of multiple users, all aiming to minimize their travel costs.

Thus, in this work, we consider pure bottleneck congestion dynamics, with spaceless vertical queues as introduced by Vickrey [1969] where, at the link level, an increase in density never corresponds to a decrease of flow. We combine this with endogenous departure time choices, utilizing thus scheduling cost functions (defined in Section 1.3). This model has received substantial attention from the transportation economic community but mainly on single-link networks ([Small, 2015, Li et al., 2020]) . Here, it is well-known that hypercongestion effects do *not* occur [Arnott et al., 1990]. Extending from a single link to multiple perfectly parallel links (i.e., a collection of completely separate routes from the origin to the destination) does not introduce any new behavior: hypercongestion still does not occur. In this work, we will consider arbitrary network topologies, which have received much less attention in the literature of hypercongestion.

We distinguish two aspects of hypercongestion that generally coincide for regular single-link models, but that become useful to distinguish in our network setting.

Speed-flow hypercongestion. We consider this form of hypercongestion to occur when a backward-bending curve manifests in a macroscopic equivalent of Figure 5.1a or

[2]For example, consider a highway with two consecutive off-ramps, with different destinations. If the later off-ramp becomes sufficiently congested, the resulting queue can block traffic exiting on the first off-ramp, resulting in a lower overall network capacity.

Figure 5.1b, using appropriate macroscopic equivalents of speed, density and flow (throughput). We will make this precise, and detail how we measure these quantities, in Section 5.2.

Generally, we would expect that the network becomes more congested as the desired arrival time approaches, after which it decreases. So this form of hypercongestion may show itself as a period of time before the desired arrival where throughput starts to *decrease*, or alternatively, a period after the desired arrival time where the throughput *increases*.

Throughput hypercongestion. This form of hypercongestion is exhibited when the imposition of optimal (first-best) pricing to decentralize the social ("system") optimal departure pattern leads to an increase in the (time-averaged) arrival flow at the destination, compared to the original no-toll equilibrium. Put differently: tolls chosen to minimize the average cost (journey time cost plus scheduling cost) lead not just to a reduction of these *average costs* (a triviality), but also to a reduction in the *generalized price*, in which the toll is also included. Thus, as phrased by [Arnott, 2013] in his discussion of the bathtub model, "[in] very congested cities optimal tolling would still benefit commuters even if the toll revenue were completely squandered!".

A reduction in the generalized price in a dynamic equilibrium with homogeneous users necessarily implies that the duration of the peak—the time between the first and last departure—decreases. Note that bottleneck models for heterogeneous users have already shown that some users can gain from the imposition of optimal tolls before revenues are recycled. But these gains do not result from an increase in rate of arrivals at the destination, and hence a shortening of peak duration, but rather from the replacement of travel delays as a dynamic equilibrium-restraint mechanism by tolls. This tends to benefit users with a high value of time, for whom paying with time is relative less attractive than paying with money [van den Berg and Verhoef, 2011]. In the present model, the reduction in travel price stems from an increase in the physical network usage and we obtain it under the, for this perspective conservative, assumption of homogeneous preferences.

Our results. We answer the main question with a resounding *yes*. We show that both forms of hypercongestion can occur, even in very small networks, and with all traffic having the same origin and destination. Our result is perhaps surprising, and seems to differ from what was generally believed. For example, Arnott et al. [2016, page 1] write on the model of bottleneck congestion:

> "While it has proved very adaptable and has generated a host of useful insights, as a model of downtown traffic congestion it is flawed since it rules out hypercongestion, assuming instead that under congested conditions aggregate traffic flow is constant."

Our primary goal is to show simply that hypercongestion *can* occur in a model with bottleneck congestion, as soon as the network structure is explicitly taken into consideration; and, that is true even if spillback congestion is ruled out. We do not examine

realistic city-sized instances or compare with empirical data. But we do show that the effect is robust, and does not require precise numerical tuning of the instance. Further, the effect can be fairly significant; for example, we can exhibit a decrease of the average generalized price using optimal tolling of approximately 20%. We examine some aspects of the sensitivity of the hypercongestion effect to various parameters in Section 5.4.

We also emphasize upfront that unlike for a single bottleneck, the two forms of hypercongestion are no longer equivalent. In particular, it is possible for speed-flow hypercongestion to occur while throughput hypercongestion does not, and vice versa.

The interesting feature of our result is that we obtain increased throughput and reduced generalized prices for finite tolls that do not fully close down certain links, like one might expect from Braess-type networks, and that satisfy necessary conditions for first-best optimal (dynamic) congestion pricing. This has important policy implications regarding the acceptability of optimal tolls.

In parallel-link networks with pure bottlenecks, the generalized price under optimal tolls is the same as for the untolled equilibrium. Our main result shows that it can strictly decrease in general. The reader may wonder whether optimal tolls can ever strictly increase the generalized price in this model. In Section 5.5, we show that this is indeed possible. Generalized prices increasing due to tolls are what commentators typically have in mind when opposing road pricing. It is noteworthy that in our model this intuitive notion may be confirmed, but may also be rejected.

Policy implications. Our findings are relevant for various pertinent policy questions in urban transport.

First, we provide a framework in which we can separate hypercongestion as arising from network interactions from hypercongestion as it may result from local, link-specific travel time functions. Insights into causes and consequences of hypercongestion are of great interest for efficient policy makings for cities around the world. Traffic congestion is without doubt one of the main challenges facing contemporary cities and hypercongestion is a particularly severe and wasteful phenomenon, representing traffic conditions for which the same flow can be reached at a higher-speed, lower-density configuration. Therefore, improving the understanding of hypercongestion helps better formulating policies to combat the most severe types of congestion.

Second, we show how optimal pricing may not only eliminate queuing but in addition decrease generalized prices before toll revenues are redistributed. This is an important question for the political and social acceptability of pricing. Among multiple reasons, the mere fact that congestion pricing normally brings societal benefits at the price of raising the generalized price for road users is often seen as a dominant cause for the very limited societal acceptability of road pricing (e.g., [Small and Verhoef, 2007]). Therefore, finding instances of road pricing that addresses the most severe type of congestion while reducing the generalized price of travel before recycling tax revenues is a particularly attractive feature.

Third, we provide insight into how lessons that can be learned from spatially homogeneous bathtub or MFD models translate into link-specific, spatially differentiated congestion pricing, as it is likely to be implemented in real applications: namely, whenever road pricing policies will not be defined to be homogeneous over space (as will be

true for the archetype bathtub model), but will instead be differentiated over key links and bottlenecks in the whole network.

Outline of the chapter. In Section 5.2 we formally define the model and the two aspects of hypercongestion that generally coincide for regular single-link models. Then, in Section 5.3, we present an instance where both forms of hypercongestion occur and in Section 5.4 we investigate further such instance. In Section 5.5 we show that, unfortunately, it is not always true that optimal tolls decrease the generalized prices but there are instances where the opposite occurs, i.e. where optimal pricing strictly increases the generalized price. Finally, in Section 5.6 we summarize our results and their implications.

5.2 Model and preliminaries

Within the economics literature, the Vickrey bottleneck model has been principally studied in the setting of single or purely parallel links. We will be particularly concerned with more complicated network structures than purely parallel links and we refer to Chapter 2 and Section 4.2 for a description of the network and equilibrium structures. We will indicate an equilibrium with (f, ℓ) and will use $z_e(\theta)$ to denote the total mass of the queue on arc e at time θ.

We consider the standard α-β-γ scheduling cost function given in Equation (1.1) with $\alpha = 2, \beta = 1, \gamma = 3$ (which imply quite standard ratios between these shadow prices α, β and γ) throughout (our results can easily be reproduced for other reasonable choices[3]).

Tolls. When pricing is implemented, a (possibly time-varying) toll constitutes a third type of disutility. It is part of the generalized *price* of travel, but since it constitutes a monetary transfer, it does not make up a societal *cost* of travel.

To discuss throughput hypercongestion, it is helpful to consider optimal tolls as a means to decentralize the societal (system-)optimum outcome. Our tolls will be charged on a per-link basis, and moreover, we will allow them to be time-varying. However, all users traversing a link at the same moment incur the same toll on the link. The only other restriction we will have is that tolls must be nonnegative. Under our assumption of inelastic overall demand, this assumption is innocent (for optimal tolls with price-sensitive demand, it will still be satisfied).

So formally, we describe the tolling scheme on a link e by a function $\mu_e : \mathbb{R} \to \mathbb{R}_+$. Then $\mu_e(\theta)$ will be the toll charged to a flow particle *entering* the link at time θ.

The essential notion of a dynamic equilibrium in the tolled setting remains unchanged, aside from the adjustment to the disutility of a user. No user should be able to change their route or departure time in a way that decreases their *generalized price*, which is the sum of the journey time cost, scheduling cost, and tolls paid by the user. All users should experience the same generalized price in a tolled equilibrium. The *generalized cost* for a user refers to the sum of journey time costs and scheduling costs only; note that this may differ between users in the presence of tolls. Optimal (first best) tolls are tolls which

[3]The value of α should be strictly larger than β, so that for a fixed departure time, a user would always want to arrive as early as possible to minimize their cost, and no detouring to postpone arrival is induced.

minimize the average generalized cost of the users, amongst all possible tolls. Note that they do not in general minimize the average generalized *price*. In Chapter 3 of this thesis we present an algorithm to compute them.

Hypercongestion. We now return to the two types of hypercongestion discussed earlier, and give precise formal definitions.

Speed-flow hypercongestion. Consider the evolution of the (untolled) equilibrium. Plot the inverse of the journey time $\ell_t(\theta) - \theta$, against $\nabla f_t(\ell_t(\theta))$, the inflow rate into t at time $\ell_t(\theta)$, for all choices of θ. The inverse journey time can be viewed as a measure of speed across the network as a whole.[4]

This plot is a macroscopic analogue of Figure 5.1d. Speed-flow hypercongestion is manifest by the appearance of a backward-bending curve in this figure; a region where the speed (inverse journey time) and the network throughput (flow rate into t) are both decreasing.[5]

Throughput hypercongestion. This is straightforward: if the generalized price at equilibrium under optimal tolls (recall this is identical for all users) is strictly smaller than the cost (including both journey time and scheduling costs) experienced by all users in the untolled equilibrium, throughput hypercongestion is exhibited, reflecting a reduction in the peak duration and, equivalently, an increase in the time-averaged (over the full peak) arrival rate.

5.3 An instance exhibiting hypercongestion

In this section we present an instance (Figure 5.2) where both forms of hypercongestion occur, and we describe the evolution of the dynamic equilibria both when no tolls are present, and when first-best pricing is applied.

Evolution without tolls. For concreteness, we will fix the free transit times, link capacities, and the total mass Q to fixed values (chosen for numerical convenience). Precisely the same qualitative behavior holds for quite a large set of parameter choices; we discuss this further in Section 5.4. Recall that we use the scheduling cost function given in Equation (1.1) with $\alpha = 2, \beta = 1, \gamma = 3$. Let $Q = 1760, T^* = 75$, and capacities and free transit times as in Figure 5.3.

[4] But note that it does not represent an averaged physical travel speed, since the actual distance traveled depends on the route the user takes through the network.

[5] One could also consider plotting some measure of density against network outflow, which would more directly correspond to the macroscopic fundamental diagram of Figure 5.1d. The main difficulty is providing a generally appropriate definition of density that can sensibly be compared to network throughput at a specific moment in time. The total number of users in the system at a particular time has little direct relationship to the inflow rate at t at the same moment in time. This is not an issue in the bathtub model, because of the assumed spatial homogeneity. While other definitions are possible, we prefer to stick to a single notion.

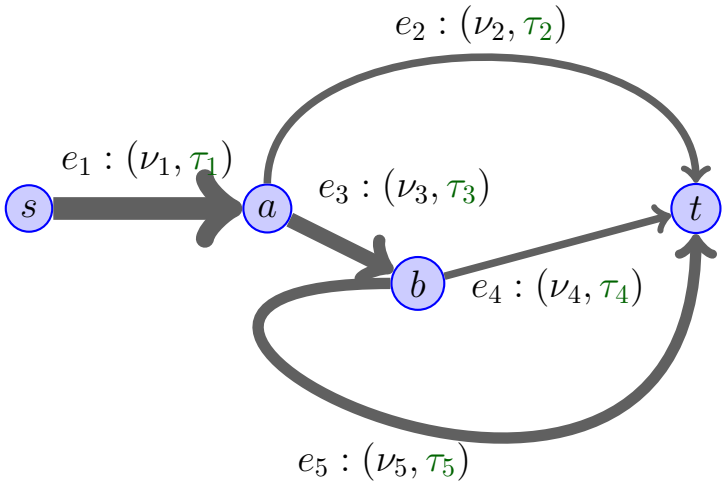


Figure 5.2: Instance where both forms of hypercongestion occur; the label of an arc represent the capacity (first element) and the free transit time (second element). Thicker arcs represent links with relative high capacity, and longer arcs represent links with relative long free transit time as assumed in the numerical example.

	e_1	e_2	e_3	e_4	e_5
Capacity (ν_e)	30	10	20	10	20
Free transit time (τ_e)	0	5	0	0	25

Figure 5.3: Remaining parameters for the instance considered in this section.

The evolution of the dynamic equilibria when no tolls are present can be described through six *phases* (see Figure 5.4). A single phase corresponds to a time interval where (1) the rate of users departing s is constant, and (2) these users all make the same aggregate route choices, i.e., the same fraction of users take any given path, for any given departure time in this interval. Within a phase, queue waiting times on any given arc change *linearly* over time. Of most interest to us will not be the rate of change of the queue length with respect to time, but *with respect to the time of departure from s*. Since a user departing from s at time θ would arrive at a vertex v at time $\ell_v(\theta)$, the rate of interest for a queue on arc vw is precisely $\frac{1}{\nu_{vw}} \frac{dz_{vw}(\ell_v(\theta))}{d\theta}$. This quantity is shown for each queue in each phase in Figure 5.4; when it is positive, the queue is growing, and when it is negative, it is depleting.

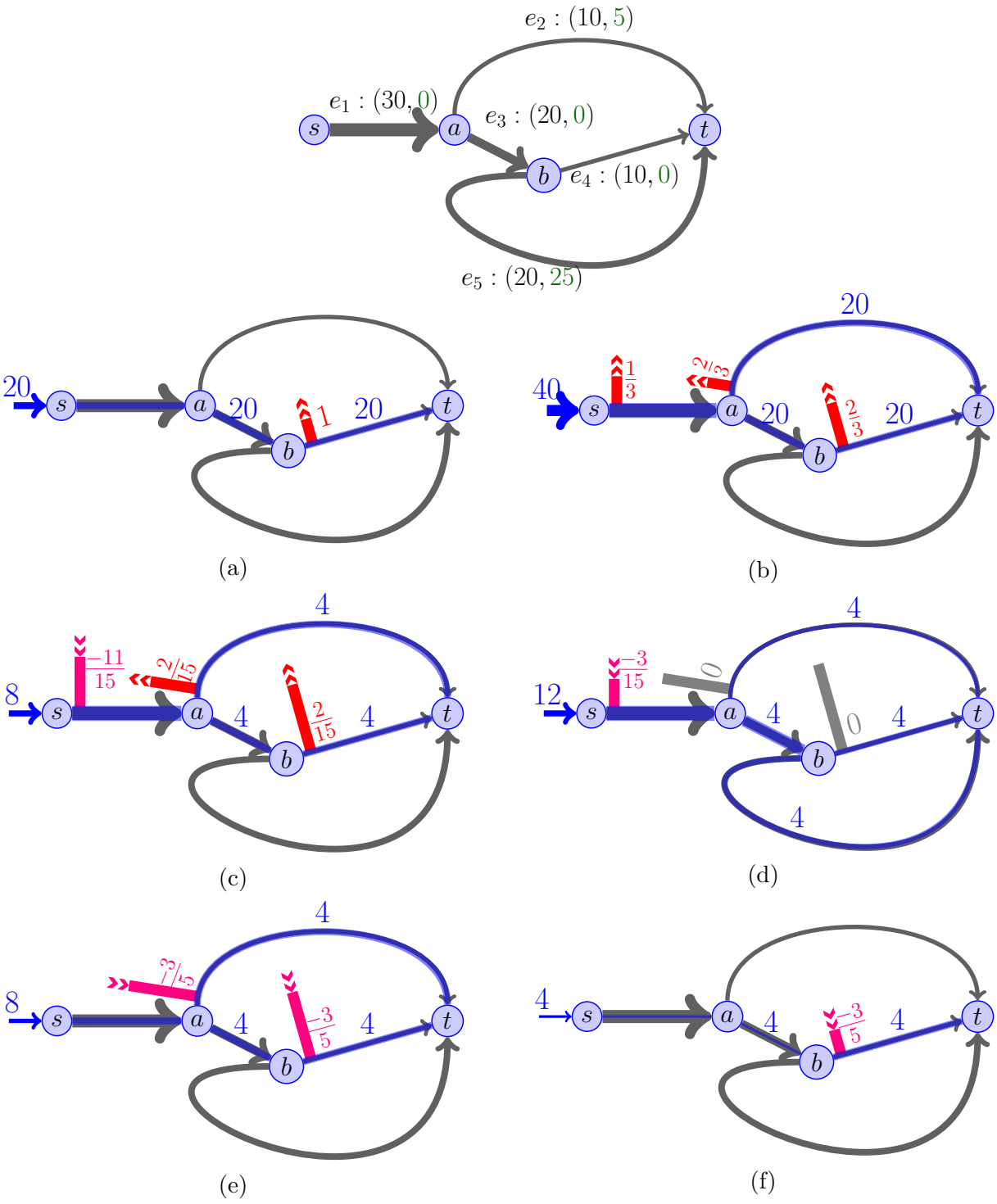


Figure 5.4: Chronological evolution without tolls: in blue the inflow into the network, the routes that the drivers take and how these drivers split among these routes; in red, purple and gray the queue waiting times that change for consecutive drivers and their rate of change: in red the ones that increase, in purple the ones that decrease and in gray the ones that stay constant.

We now describe the six phases of the equilibrium.

(a). The first phase starts when the network is empty and the first agent departs, at time 6, taking the unique shortest path $e_1e_3e_4$. Per unit of time, 20 drivers take this path and hence a queue grows on e_4 at rate 1: this means that a particle of this phase appearing at s an ε amount of time later than the first one, experiences a queue waiting time of ε on e_4. Notice that the cost of these two particle is the same since the latter pays 2ε more in journey delay cost and, arriving 2ε time later, pays 2ε less in scheduling cost.

(b). At time 11, when the queue waiting time on e_4 equals the free transit time of e_2, the situation changes. If phase 1 were to continue, the travel time on e_4 would continue to increase, meaning that e_1e_2 would be a strictly shorter route from a to t. Instead, users start to take the path e_1e_2 in addition to the path $e_1e_3e_4$. At this point the inflow into the network also increases, from 20 to 40, with the effect that queues simultaneously grow on e_1, e_2 and e_4. Notice that, along these two paths, the total queue waiting time grows at rate 1 and hence all the particles of this phase incur the same cost, equal to that in phase (a).

(c). The situation changes again at time 40.5. The agent that departs at this moment arrives at the destination at time T^*. Since the penalty for being late is nonincreasing over (arrival) time, the journey time has to decrease swiftly for agents departing from this moment onward. As a consequence, the inflow into the network drops to 8. In this phase the queue on e_1 starts to deplete while the queues on e_2 and e_4 continue to grow. This happens since the capacity (and hence the outflow) of the arc e_1 is bigger than the capacity of the arc e_2 and e_4 combined.

From this phase onward, the total queue waiting time decreases at rate $\frac{3}{5}$: a particle departing an ε amount of time later pays $\frac{3}{5}\alpha\varepsilon = \frac{6}{5}\varepsilon$ less in journey delays and $\gamma \cdot \frac{2}{5}\varepsilon = \frac{6}{5}\varepsilon$ more in scheduling cost, keeping average cost constant as required for equilibrium.

The instance has been chosen so that due to the growing queue on e_4, the travel time on e_4 eventually equals the free transit time on e_5, at which point the phase ends.

(d). The next, fourth, phase is crucial and starts at time 43. The path $e_1e_3e_5$ starts being utilized, alongside all the paths that were used in the previous phase and thus the inflow rate increases to 12. All the drivers of this phase will split evenly among the three routes. At this point the queue lengths on e_2 and e_4 stay constant while the queue on e_1 continues to deplete. Note that this phase will cause an inflow rate into t particularly high; equal to 30, the capacity of e_1. The phase ends once the queue on e_1 empties.

(e) and (f). In the final two phases, all remaining queues gradually deplete, and route choices consolidate to the shorter paths. Phase (e) starts at time $56 + \frac{1}{3}$ and phase (f) at $89 + \frac{2}{3}$. The final agent to depart arrives at t just at the moment when the queue on e_4 depletes, at time 98, thus experiencing again an empty network.

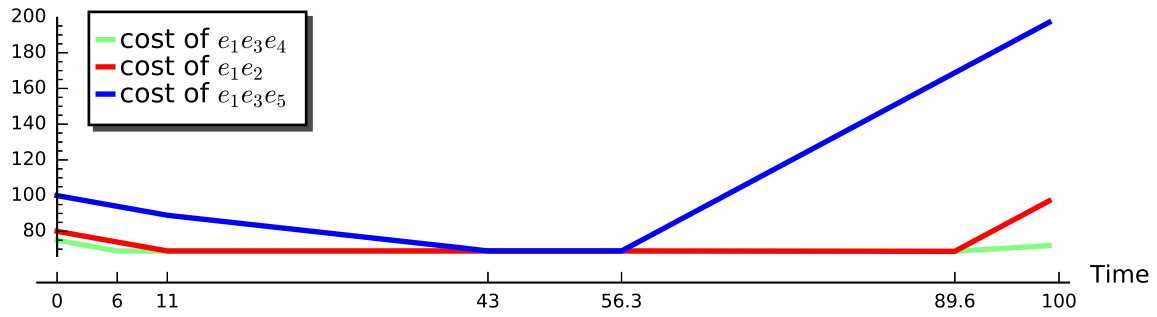
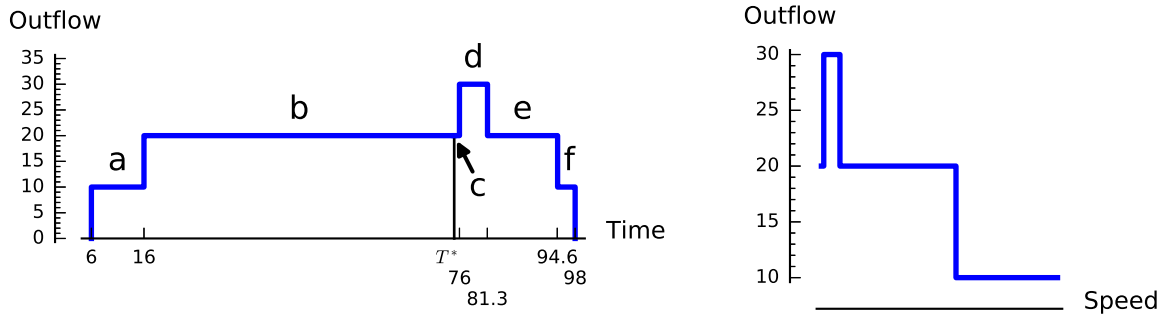


Figure 5.5: Commuting cost of the different routes under the no-toll equilibrium. The horizontal axis represents the departure time. A route is active when its cost is equal to the minimum.

Figure 5.5, displaying how average cost evolves over time for routes when they are and when they are not used, confirms that the patterns described jointly provide a dynamic equilibrium, in departure time *and* route choice.

The outflow over time is shown in figure Figure 5.6a. After T^* the outflow of the network does not behave monotonically: it first increases and then decreases. Since the schedule delay cost increases over time, after T^* the journey time of the agents arriving at t decreases over time. Hence, after T^*, we have an increase in the outflow with an increase of speed (inverse journey time), as shown in Figure 5.6b. Therefore, the instance exhibits speed-flow hypercongestion.

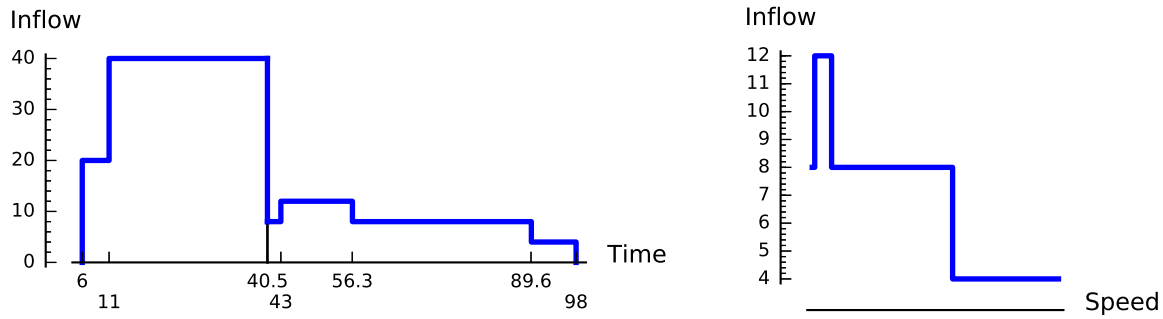


(a) (b)

Figure 5.6: On the left the plot showing the outflow of the network in any given time; the letters indicates the phase to which the outflow is associated to. On the right the plot indicating the relation between the outflow of the network and the speed (inverse journey time) for the agents arriving after T^*.

This result does not rely on our choice of measuring the passing flow at the sink of the network; it is in fact quite robust. For instance, we see exactly the same behavior if we compare the *inflow* into the network at a given moment with the speed (inverse journey time) of the agents departing at the same moment. The inflow over time is shown in Figure 5.7a; after time 40.5 (which is the departure time of the agent arriving at

destination at time T^*), we again see that there is a moment where the inflow increases. Since the journey time (meaning the inverse speed) of agents departing from s decreases from time 40.5, we again observe an increase in the inflow with an increase of speed (Figure 5.7b).



(a) (b)

Figure 5.7: On the left the plot showing the inflow of the network in any given time; the letters indicates the phase to which the outflow is associated to. On the right the plot indicating the relation between the inflow of the network and the speed (inverse journey time) for the agents departing after time 40.5.

Remark. The textbook exposition of the backward-bending speed-flow relation (Figure 5.1a) assigns at most one flow level to each positive speed. This is not the case for the effective speed-flow relation in our example. Speeds (inverse journey times) within an appropriate interval are seen twice; once before the desired arrival time and once after. For some speeds, the arrival rates at these two corresponding moments differ. This does not occur in the setting of a single link, or multiple parallel links, and represents another noteworthy feature of general networks from the conceptual viewpoint. Moreover, this could be one reason—among others—that empirical plots of flow versus speed display so much scatter.

Evolution with optimal tolls. Under optimal pricing all three paths from s to t are used, and no queue is ever formed. An example of optimal pricing is obtained by tolling one arc per path and by increasing these tolls linearly at rate β until time T^* and then decreasing it at rate $-\gamma$ (see Figure 5.8). This way each tolled arc is in a unique path, and that path is used continuously in the time interval where its arc is tolled. Chronologically, the evolution can be described through six phases. Each phase indicates a time interval in which the particles arriving at t take the same paths. The last three are mirrored but scaled version of the first three: the scaling is due to the ratio of β and γ. In the first and last phase (Figures 5.9a and 5.9f) agents arrive at t only from the path $e_1e_3e_4$; in the second and fifth phase (Figures 5.9b and 5.9e) they come from paths e_1e_2 and $e_1e_3e_4$ and in the third and fourth (Figures 5.9c and 5.9d) they arrive from all the paths e_1e_2, $e_1e_3e_4$ and $e_1e_3e_5$.

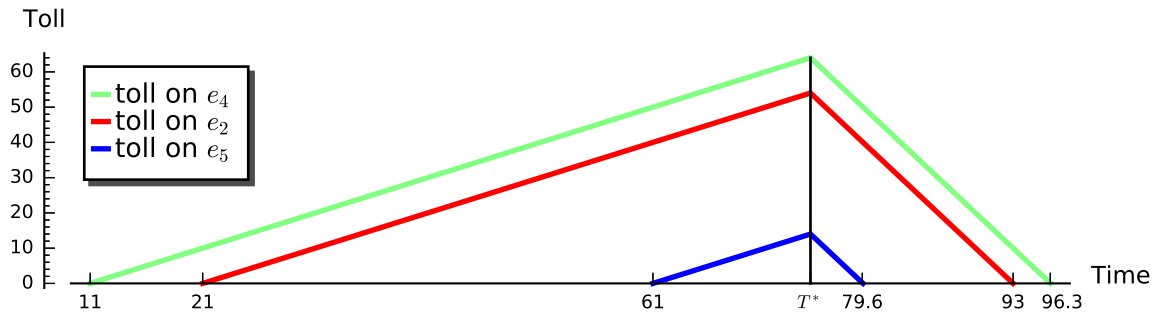


Figure 5.8: Example of optimal tolls for the studied instance.

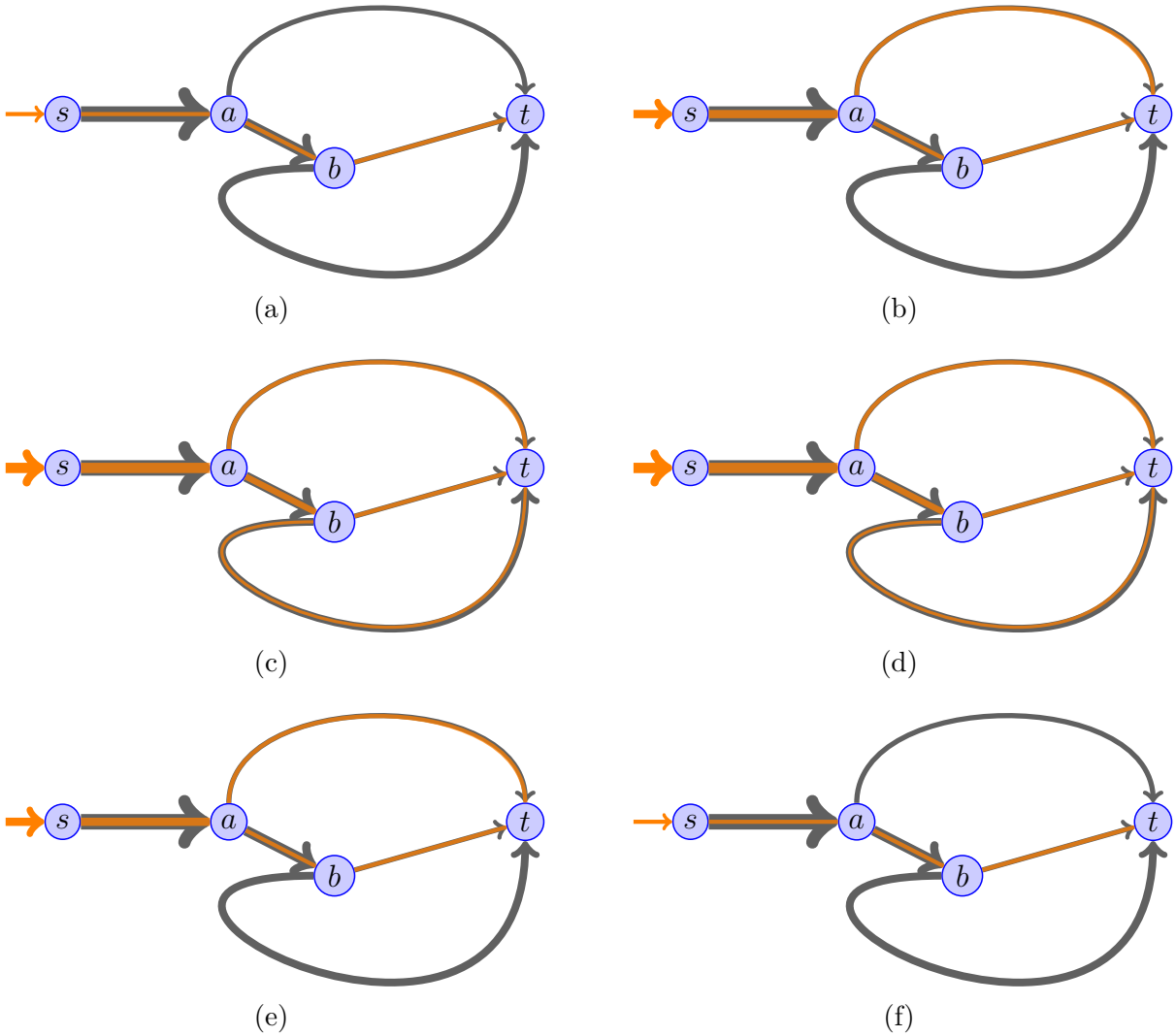


Figure 5.9: Chronological evolution with tolls: in orange the routes that the agents arriving at t have taken.

In both the equilibria the first agent that arrives at t takes the same path $e_1e_3e_4$ and incurs no cost in tolls, but the arrival time is later when first-best pricing is applied: 6 under no toll and 11 under optimal pricing (Figure 5.10). This implies that his general price, and hence that the general price of all the agents, is lower under first-best pricing. Similarly, the very last driver arrives at 98 under no toll and $96 + \frac{1}{3}$ under optimal pricing (Figure 5.10). With optimal tolls, the peak duration is therefore shorter and, as a consequence, the time-averaged throughput is higher. This occurs because the path $e_1e_3e_5$ is active over a longer period with optimal tolling than without. Thus, the instance also exhibits throughput hypercongestion: the time averaged arrival flow is higher, the arrival window shorter and the generalized price is lower in the optimum compared to the no-toll equilibrium.

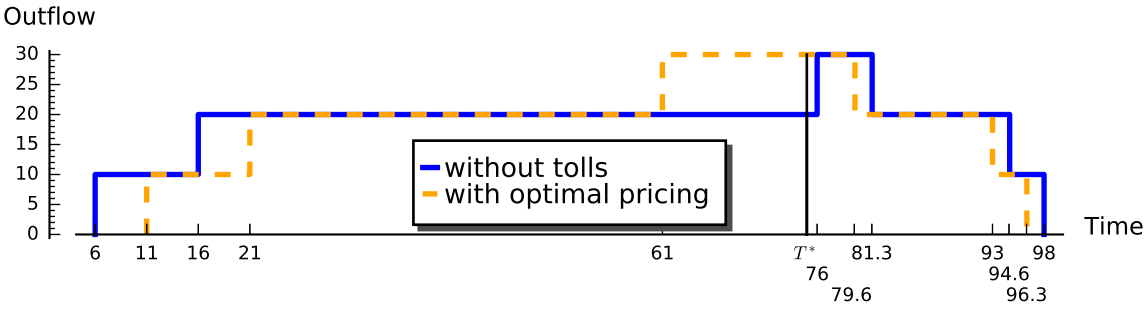


Figure 5.10: Outflow over time of the no-toll and of the best-pricing equilibria.

5.4 Sensitivity analysis

Consider again the network of Figure 5.2, that we reproduce here for convenience:

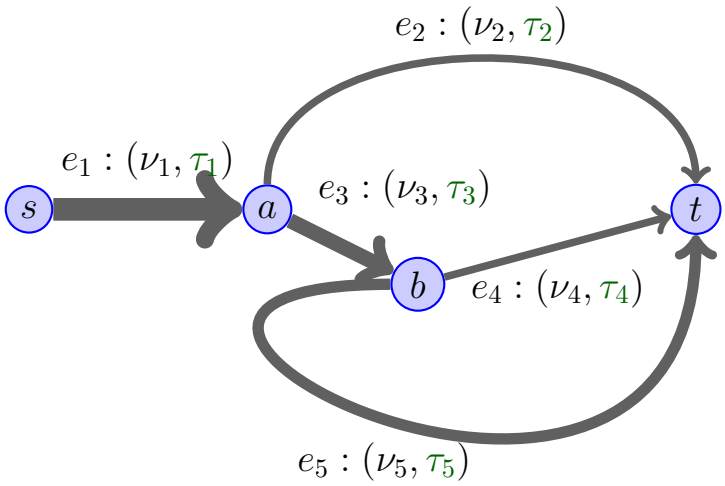


Figure 5.11: Instance where both forms of hypercongestion occur.

The following theorem specifies a set of parameter choices for which both forms of hypercongestion occurs in this network.

Theorem 5.1. *Suppose all the following conditions are satisfied, for an instance of the network shape shown in Figure 5.11:*

$$\tau_3 + \tau_4 < \tau_2 < \tau_5 \tag{5.1}$$
$$\tfrac{\alpha}{\alpha-\beta} \cdot \nu_4 \leq \min\{\nu_1, \nu_3\} \tag{5.2}$$
$$\nu_3 \geq \max\{\nu_1 - \nu_2, \tfrac{\alpha}{\alpha+\gamma}\nu_4\} \tag{5.3}$$
$$\nu_5 \geq \nu_3 - \nu_4 \tag{5.4}$$
$$\nu_2 + \nu_4 < \nu_1 < \tfrac{\alpha}{\alpha-\beta} \cdot (\nu_2 + \nu_4). \tag{5.5}$$

Then there exists a nonempty interval[6] *of total demand for which both speed-flow and throughput hypercongestion occur.*

Notice that constraint (5.1) says that $e_1e_3e_4$ is the shortest path and $e_1e_3e_5$ is the longest one; constraint (5.2) says that e_4 has relative small capacity while constraints (5.3) and (5.4) say that e_3 and e_5 cannot have a capacity that is too small; finally, constraint (5.5) gives a lower and an upper bound to the capacity of e_1.

These constraints ensure that the constructed equilibria (both untolled and tolled) have a sequence of phases analogous to the ones considered in the previous section (see Figures 5.4 and 5.9). Analogous in the sense that, within a phase, users take the same paths and queue grow and deplete on the same arcs, even though the rate of change of the queues and the length of the phases differ. We omit an explicit proof of the theorem since it's very computational.

Theorem 5.1 implies that after fixing the graph structure for the instance, there is a range of parameters that produce the two forms of hypercongestion, confirming that the occurrence is not a peculiarity that requires a very specific, fine-tuned combination of parameters.

From an efficiency viewpoint, throughput hypercongestion most clearly reflects the inefficiency involved. The fact that time-averaged throughput is in fact higher in the optimum, and the generalized price consequently lower, adds a clear second "dividend" to the implementation of pricing, on top of the well-known favorable impact on queues. It also suggests that social support for pricing might be higher than often thought, as also without recycling of tax revenues and also if bottleneck capacities exhibit no "drop" due to queuing, users may already benefit from the imposition of optimal tolls. It is therefore of interest to see how this type of hypercongestion varies with some key parameters of the model.

Figure 5.12 shows the ratio, for the instance of Figure 5.3, between the generalized price of the equilibrium without tolls and the generalized price of the equilibrium with optimal pricing as a function of the total users demand, using $\alpha = 2, \beta = 1, \gamma = 3$. At first, the ratio is equal to 1. This scenario reproduces the behavior on a single bottleneck link, since all the agents takes only the shortest path. As the total demand increases beyond a critical

[6] Such an interval is (I_a, I_b) with $I_a = (\tau_2 - \tau_3 - \tau_4)\alpha\nu_4(\frac{1}{\beta} + \frac{1}{\gamma}) + (\tau_3 + \tau_5 - \tau_2)(\nu_2 + \nu_4) \cdot (\frac{\alpha}{\gamma} + \frac{\nu_1}{\nu_1 - \nu_2 - \nu_4})$ and $I_b = I_a + (\tau_3 + \tau_5 - \tau_2)\frac{\nu_1}{\gamma} \cdot \frac{\alpha\nu_2 + \alpha\nu_4 - \alpha\nu_1 + \beta\nu_1}{\nu_1 - \nu_2 - \nu_4}$.

value of around 1200, the generalized prices ratio first increases and then decreases, but it always remains strictly above 1. The red region demarcates the boundary of the range of choices for the total demand where speed-flow hypercongestion occurs. This shows that throughput hypercongestion can occur while speed-flow hypercongestion does not.

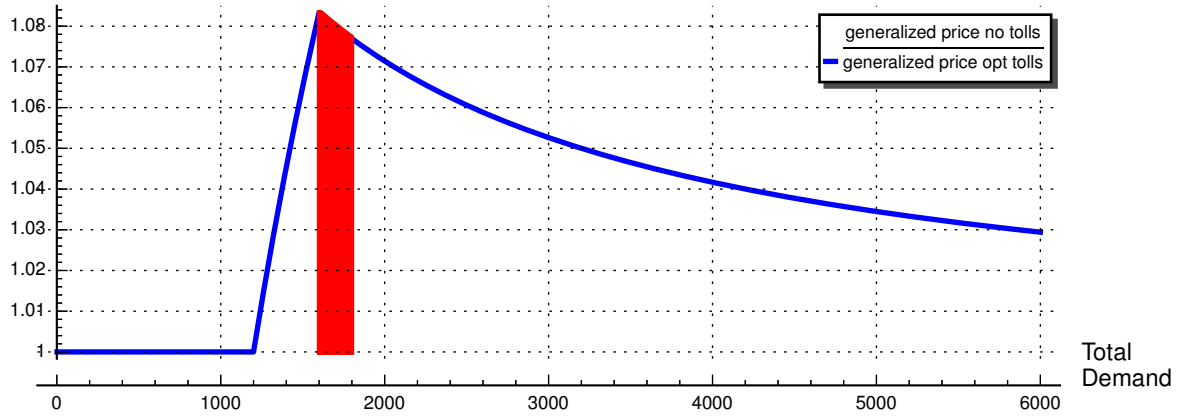
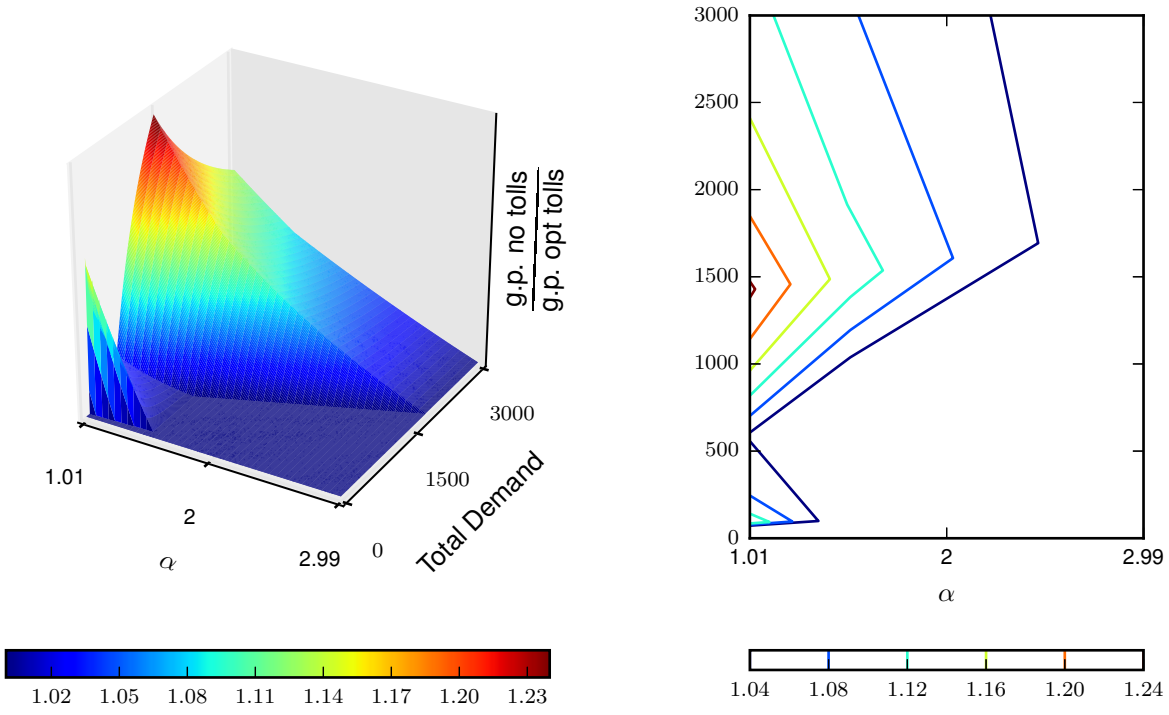


Figure 5.12: Ratio between the generalized price of the equilibrium without tolls and the generalized price of the equilibrium with optimal pricing.



(a) (b)

Figure 5.13: The ratio between the generalized prices of the equilibrium without tolls and the equilibrium with optimal pricing as a function of α and the total demand; (b) shows a contour plot.

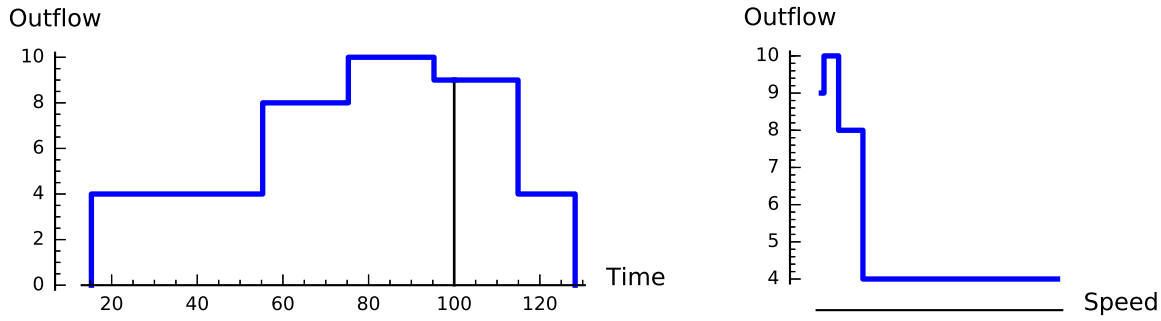
Figure 5.13 shows, again referring to the instance of Figure 5.3, the ratio between the generalized price of the equilibrium without tolls and the generalized price of the equilibrium with optimal pricing as a function of α and the total demand. Here the value of α varies in $(\beta, \gamma) = (1, 3)$. The ratio tends to increase when α becomes smaller, which is intuitive as a lower α would lead, ceteris paribus, to longer equilibrium queues. Given α, we observe a maximum ratio for some intermediate value of total demand, just as in Figure 5.12.

Separating the two forms of hypercongestion, and hypercongestion before the peak. As already noted, the previous example shows that throughput hypercongestion can occur without the presence of speed-flow hypercongestion. The opposite can also happen: speed-flow hypercongestion can occur while throughput hypercongestion does not. An example is obtained on the network of Figure 5.11 with $Q = 750$, $T^* = 100$ and capacities and free transit times shown in Figure 5.14.

	e_1	e_2	e_3	e_4	e_5
Capacity (ν_e)	9	4	6	4	4
Free transit time (τ_e)	0	20	0	0	20

Figure 5.14: Further parameters for an instance exhibiting speed-flow, but not throughput, hypercongestion.

Figures 5.15a and 5.15b shows that the instance exhibits speed-flow hypercongestion: there is a period of time before the desired arrival time where the throughput in the equilibrium decreases. Since, at this point, the journey time of the agents arriving at t increases over time, we have a decrease in the outflow with a decrease of speed (inverse journey time). Figure 5.16 shows that the instance does not exhibits throughput hypercongestion.



(a) (b)

Figure 5.15: Speed-flow hypercongestion of the instance described in Figure 5.14: On the left the outflow in any given time; On the right the relation between the outflow and the speed (inverse journey time) for the agents arriving before T^*.

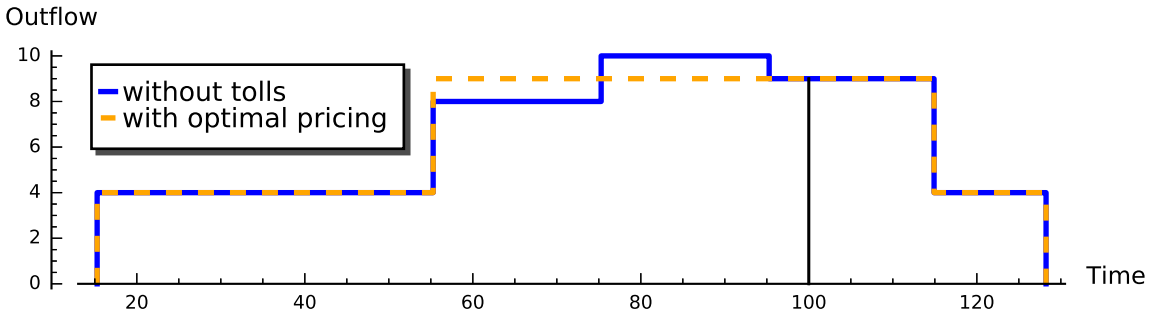


Figure 5.16: Outflow over time of the equilibrium of the instance described in Figure 5.14, both without tolls and under optimal pricing.

This last example also demonstrates that speed-flow hypercongestion can occur *before* the peak (recall that our main example in Section 5.3 exhibited speed-flow hypercongestion after the peak). Since in the empirical literature on hypercongestion at the network level (e.g., [Geroliminis and Daganzo, 2008, Daganzo et al., 2011]), the effects do not seem to be restricted to occurring only after (or only before) the peak, this is important.

Simpler network topologies. The examples discussed so far are not the smallest ones to exhibit hypercongestion. The following instance (Figure 5.17) with three links also exhibits both forms of hypercongestion, with the usual choice of α, β, γ along with $Q = 400$. We have chosen $T^* = 15$.

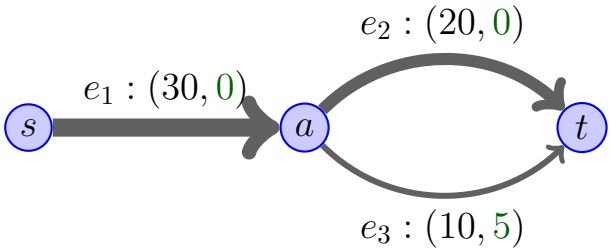
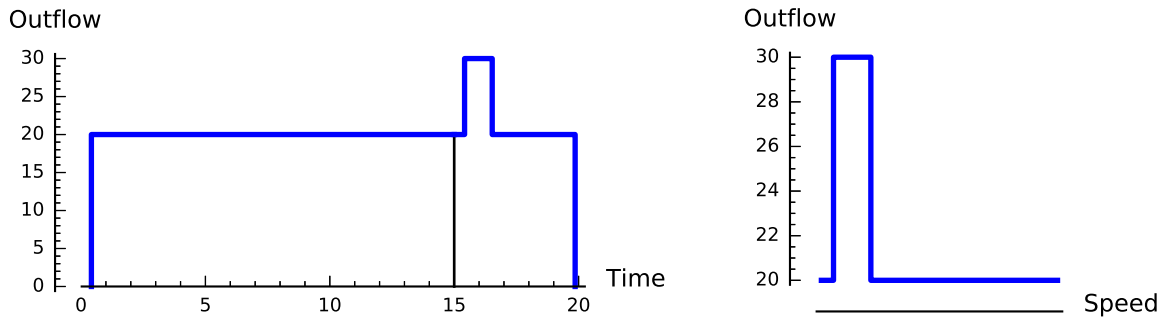


Figure 5.17: The label of an arc represent the capacity (first element) and the free transit time (second element).

Figures 5.18a and 5.18b shows that the instance exhibits speed-flow hypercongestion, and Figure 5.19 that it exhibits throughput hypercongestion.



(a) (b)

Figure 5.18: Speed-flow hypercongestion of the instance described in Figure 5.17: On the left the outflow of the network in any given time; On the right the relation between the outflow of the network and the speed (inverse journey time) for the agents arriving after T^*.

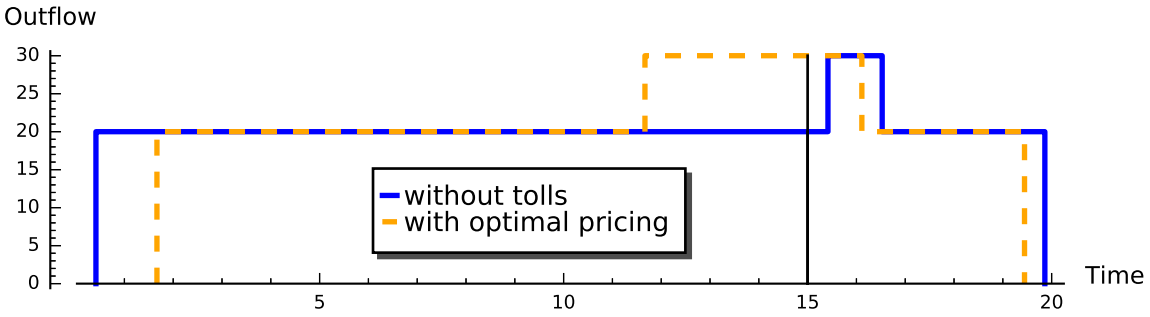


Figure 5.19: Outflow over time of the equilibrium of instance described in Figure 5.17, both without tolls and under optimal pricing.

The fact that hypercongestion occurs for such a simple network topology provides further evidence that hypercongestion is not rare for this type of model.

We remark that we have focused on the slightly larger five-arc instance in order to exhibit some phenomena beyond the existence of the two forms of hypercongestion. In particular, our observations that (i) that the two forms of hypercongestion are distinct: either form can occur without the other; and (ii) that speed-flow hypercongestion can occur *before* the peak rely on our larger example.

5.5 First-best pricing can strictly increase the generalized price

The fact the optimal tolling decreases the generalized price is remarkable, and runs counter intuition. It also is at odds with results from other dynamic models of traffic congestion, including the conventional single-link bottleneck model, where the generalized price does

not change when an optimal toll is implemented, and models of flow congestion such as the one proposed by Chu [1995], where the generalized price would increase. That raises the question of whether, for other parameter combinations or networks, the current model could also produce instances where the generalized price rises instead of falls.

In this section we show through an illustration that first-best pricing does not always decrease the generalized price, and can in fact cause it to strictly increase. For this we consider the network of Figure 5.20. Figure 5.21 shows the ratio between the generalized price of the equilibrium without tolls and the generalized price of the equilibrium with optimal pricing, as a function of the total users demand (using $\alpha = 2, \beta = 1, \gamma = 3$ as before). At first, the ratio is equal to 1. This occurs since all the agents takes only the shortest path and hence we have the same behavior we would have on a single bottleneck link. As the total demand increases, the generalized price ratio first decreases sharply, and then increases asymptotically towards 1.

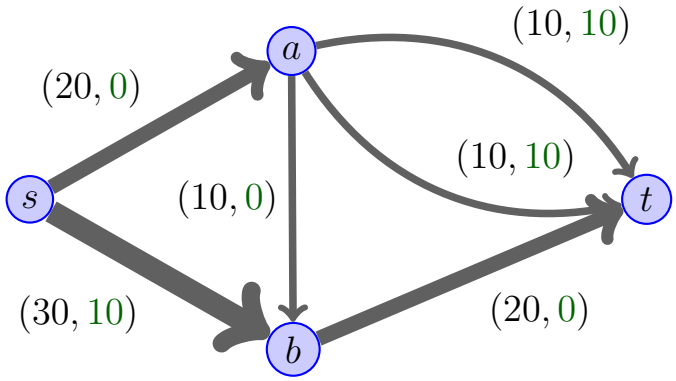


Figure 5.20: Instance where optimal tolls increase the generalized price.

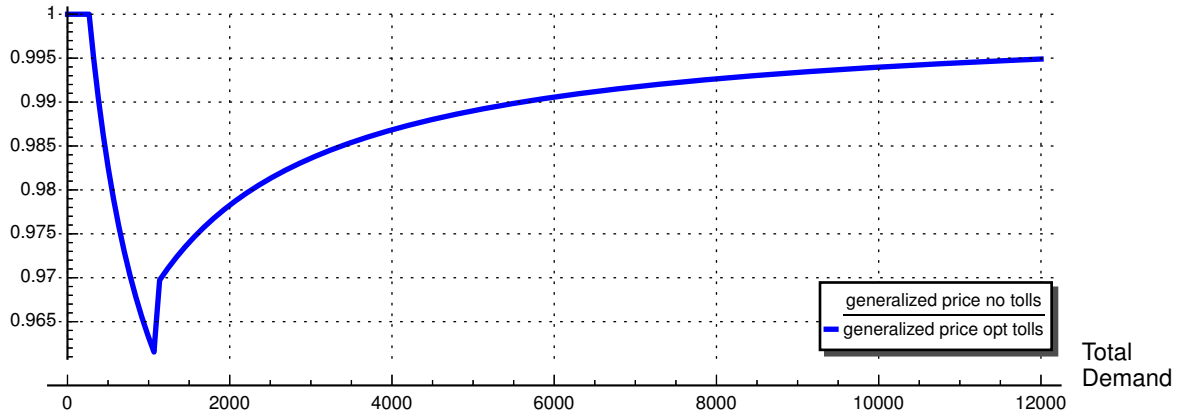


Figure 5.21: Ratio between the generalized price of the equilibrium without tolls and the generalized price of the equilibrium with optimal pricing.

5.6 Conclusion

In this chapter we showed that hypercongestion can occur as a purely emergent effect of the network interaction of multiple selfish users, all aiming to minimize their travel costs. For this we considered link congestion dynamics that do not exhibit hypercongestion and do not produce spillback effects (we considered Vickrey bottlenecks with spaceless vertical queues). In this work we used the model and algorithm defined in Chapters 3 and 4.

We distinguished two instructive aspects of hypercongestion that generally coincide for regular single-link models. One is related to the "macroscopic" versions of the fundamental traffic diagram and the other is exhibited when the imposition of optimal (first-best) pricing leads to an increase in the (time-averaged) arrival flow at the destination and therefore to a reduction in the generalized price, i.e., the total cost that each agent pays.

These findings are of interest because they show how the same flow can be reached at a higher-speed, lower-density configurations, thus eliminating queuing. Moreover, the second form of hypercongestion that we considered shows how optimal pricing may not only eliminate queuing but in addition decrease generalized prices before toll revenues are redistributed. As opposed to the bottleneck models for heterogeneous users where some users can gain from the imposition of optimal pricing to the detriment of other users' costs, in this chapter we assumed homogeneous users and thus the reduction in travel price comes from an increase in the physical network usage. This is an important result for the political and social acceptability of pricing. However, we also showed that this is not always the case and that there are instances where optimal pricing can increase the generalized price. Finally, our pricing policy is not defined to be homogeneous over space, as in the case of other models that study hypercongestion, but it is differentiated over key links and bottlenecks in the whole network, as it is likely to be implemented in real applications.

Fully understanding the causes of hypercongestion as empirically observed remains a challenging task. It is unclear how the form of the network—for instance, it's topology—effects the prevalence of hypercongestion. It is well know that it does not occur in parallel link networks and our construction shows that it may occur in very simple networks, but little is known beyond that. It would be interesting to investigate what happens in the Vickrey bottleneck model on instances more indicative of real city networks, and with multiple origin-destination pairs. It is also natural to ask what impact other specifics of the model have on hypercongestion. For example, one can consider user heterogeneity, in the value of time or in scheduling preferences; or models where there is uncertainty in the delays experienced on links.

Chapter 6

Revisiting the corridor problem with discrete bottlenecks

The results in this chapter are being prepared for publication and were obtained thanks to the collaboration of Neil Olver and Erik Verhoef.

6.1 Introduction

The *corridor model*, in its original form, introduced by Arnott [2001] and considered in detail by Arnott and DePalma [2011b], is a model of traffic congestion on a highway entering a city during the morning rush hour. Their model is continuous in both space and time; a residential density along the highway is prescribed, from which the total amount of traffic that will enter the system within any interval can be determined. All users wish to travel to the CBD, which is located at the end of the highway. (Here, the morning rush hour is being modeled; one can of course consider the evening rush hour as well, in which case users will all depart the CBD, and leave at various locations along the corridor.) Users can choose their departure time, and suffer the usual α-β-γ costs; the journey time cost is α, and there is a scheduling cost of γt incurred by arriving t time units late, or βt for arriving t units early. The kinematic LWR model is then used to describe the congestion on the corridor. That is, there is a partial differential equation involving flow and density on the corridor through space and time, with a flow-speed relationship prescribed.

The resulting model turns out to be challenging, and unsatisfactory in some key respects. In particular, Arnott and DePalma find that in general, an equilibrium need not exist in their model. In fact, they showed that unless the residential density satisfies certain conditions (that the residential density is zero in some neighborhood of the CBD), no equilibrium is possible; and they were not able to derive sufficient conditions for existence. They were able to give some interesting insights on the qualitative behavior of an equilibrium in their model (if it exists); but a satisfactory treatment of the model seems out of reach. The LWR model alone, involving a partial differential equation, is already very difficult to handle. Coupling this with equilibrium conditions of the corridor model appears to make things dramatically more difficult.

Given the complexity inherent in the LWR model, a natural direction is to consider instead a simpler model of congestion. Consider, instead, a highway modeled as a sequence of many Vickrey bottlenecks with vertical (spaceless) queues. The nodes between consecutive bottlenecks will represent discrete entry points into the corridor; we will call these *on-ramps*. Instead of a residential density distribution, we have a residential demand vector indicating the total mass of users that wish to enter on each on-ramp. In two respects, the model we consider goes beyond the original Arnott-DePalma corridor model:

- We allow for arbitrary capacities on the corridor links.
- We allow for fairly general scheduling cost functions (which describe the disutility experienced by a user arriving at the CBD at a certain time, in addition to their journey time cost). Essentially any convex, piecewise linear function may be considered.

This model, essentially, was introduced and discussed in some detail by Akamatsu et al. [2015] (we will refer to their paper as *AWH* in what follows). However, while they introduce their model in continuous time, all their analysis is actually on a variation of the model which is discrete not only in space (in terms of a finite number of on-ramps) but discrete *time* as well. They consider this discrete-time variant primarily as an approximation to the continuous time model that is more amenable to their techniques. Without defining this in detail, we will call this the *discrete time Vickrey corridor*, to distinguish with the *continuous time Vickrey corridor* that we will consider in this chapter. A further feature that they consider in their model, which we will not, is a form of user heterogeneity. All users have the same "base" preferences (in terms of desired arrival time, per-unit journey time cost, and scheduling cost function). However, they allow for (and for some results, require) that the actual disutility experienced by a user is not just the base disutility, but the sum of this with an additional random component. This random perturbation to a user's scheduling cost is drawn from a fixed distribution; since there are a continuum of users, this means that a given perturbation will be experienced by a fixed density or proportion of users at any given location. Some of their results require that this distribution admits a density (ruling out the homogeneous case).

Our results. We now summarize the main results of our work, and compare in particular with AWH. We will only consider the morning rush hour (AWH consider both the morning and evening rush hour). There is no particular difficulty with the evening rush hour, which could also be dealt with using our approach (and in fact is easier in certain respects). But we prefer to focus our discussion on the (more prevalent) morning peak.

- *Existence.* AWH do prove existence of equilibria for the discrete time Vickrey corridor. Their result applies for both homogeneous and heterogeneous users (in the restricted sense discussed above), and for morning and evening rush hour.

  We prove existence in the continuous time Vickrey model (with homogeneous users). While AWH somewhat suggest that their discretization is to avoid "unnecessary mathematical complications", it is not that clear to us whether existence results for the continuous model can be derived from their approach to the discrete model.

- *Uniqueness.* AWH also consider uniqueness, and obtain some significant results in this direction. However, they are *not* able to show uniqueness of equilibria in the discrete time Vickrey corridor with homogeneous users. They require heterogeneity, in the form discussed above: the utility of a user is subject to a random perturbation, drawn from a distribution that admits a density.

  While the heterogeneity assumption that AWH make can certainly be motivated, we think that the model we consider—homogeneous users, continuous time—is the most basic and natural one, so we believe that uniqueness for this model is of significant interest.

- *Structural and algorithmic insights.* Perhaps the most important contribution of this chapter is that our perspective gives a rather clear perspective on the structure of equilibria. We can give a qualitative description of how equilibria behave.

  We also provide an efficient algorithm for computing an equilibrium. AWH give an algorithm for the discrete case based on solving *linear complementarity problems* (LCPs). These are not known (and not believed) to be efficiently solvable in a theoretical sense (in particular, the problem of computing an equilibrium of a bimatrix game can be formulated as an LCP), though in practice they are relatively tractable. The LCPs they require solving, grow in size depending on the number of time steps in the instance, however. We give an efficient (polynomial time) algorithm for the continuous-time version. It's running time is $O(n(n+\omega))$, where ω is the number of breakpoints in the scheduling function. In practice, for moderately large corridor sizes we expect the algorithm to be extremely fast. So our algorithm results seem more satisfactory, both from a theoretical standpoint and a practical one.

Our approach. We cleanly separate two difficulties by developing first a model with *sensitive* demand, where each user have some fixed disutility that they are willing to suffer, and otherwise will not travel. This disutility threshold is the same for all users at a given location on the corridor, but may differ between locations. This turns out to be substantially easier than the case of insensitive demand at each location, and the structural results are derived in this context. In particular, we observe that the sensitive demand model is completely equivalent to a situation where all users have not only the same destination, but the same origin, but in an augmented network. This augmented network does have multiple possible routes between origin and destination however, so we trade the complexity of multiple origins with the complexity of route choice.

Fortunately, we can build on existing insights into this single-origin single-destination model in general networks. This stream of literature has its roots in the theory of *flows over time* in combinatorial optimization, which concerns questions like sending as much flow as possible from an origin to a destination in a network before a given deadline (see Section 1.2 and Chapter 2).

We remark that there is a connection with the approach AWH take in terms of using what they call a "Lagrangian-like coordinate system". Essentially, this means that the main measure of "time" at some location in the corridor is not the real "wall clock" time at that location, but the *time a user would arrive at the CBD in the equilibrium if departing*

at the given moment. This is indeed exactly the viewpoint taken in the above stream of literature that our approach draws on.

After obtaining detailed results for the sensitive model, we elucidate the relationship between the sensitive and insensitive demand models. We show that there is a one-to-one map between demand vectors in the insensitive model, and disutility thresholds in the sensitive model. This allows us to deduce existence and uniqueness in the insensitive model as a consequence of these for the sensitive model. It also allows us to import qualitative observations of the equilibrium structure from the sensitive model to the insensitive model.

Outline. We begin by formally introducing the model in Section 6.2. We introduce not only the insensitive Vickrey corridor model that is our main motivation, but also the variants we will use in order to pursue our analysis; the sensitive version.

In Section 6.3, we develop an understanding of equilibria in the sensitive model. As well as demonstrating existence and uniqueness, we will obtain a detailed understanding of the structure of equilibria. This will be crucial for our main results on the insensitive model (including existence and uniqueness), which are obtained in Section 6.4.

Finally, in Section 6.5 we show some examples of equilibrium behavior, and discuss our structural insights at a high level.

6.2 Model and preliminaries

Some notation. We write $[z]^+$ for the nonnegative part of z, i.e., $[z]^+ = \max\{0, z\}$. $\mathbb{R}_+$ denotes the nonnegative real numbers, and $\mathbb{R}_{++}$ the strictly positive real numbers. For a positive integer n, we use $[n]$ as shorthand for the set $\{1, 2, \dots, n\}$.

A function $g : \mathbb{R} \to \mathbb{R}$ has a property *almost everywhere* if there is a function $\tilde{g}$ satisfying the property that differs from g on a set of measure zero. A function g is *piecewise constant* on a compact interval I if I can be partitioned into a finite collection of intervals such that g is constant on each interval. A function is piecewise constant (on $\mathbb{R}$) if it is piecewise constant on every compact interval. Similar definitions apply for *piecewise linear*.

6.2.1 The insensitive demand model

Our corridor is described by a directed graph with vertex set $1, 2, \dots, n, n+1 = t$, where node t represents the CBD, with arcs $(i, i+1)$ for each $i \in [n]$. We will always envision the CBD as being to the right, and nodes $1, 2, \dots, n+1$ located from left to right in that order. Arc $(i, i+1)$ is a Vickrey bottleneck with free-flow travel time $\tau_{(i,i+1)}$ and capacity $\nu_{(i,i+1)}$, which we abbreviate to ν_i for convenience. We also make the definition $\nu_0 := 0$.

There is a given mass Q_i of users at location i, for $i \in [n]$, all desiring to reach the CBD t. (We don't assign any mass to t; this would be superfluous.) Each (infinitesimal) user experiences a per-unit journey time cost α for time spent commuting. In addition, they experience a scheduling cost that is a function of their arrival time at the CBD; let ρ denote the scheduling cost function. We will require that this is piecewise linear and convex, with $\rho'(\theta) > -\alpha$ for all θ. Also ρ should satisfy $\lim_{\theta\to-\infty} \rho(\theta) = \lim_{\theta\to\infty} \rho(\theta) = \infty$; this also implies that ρ is bounded from below. We make the final assumption that ρ

has a unique minimizer, which we denote by T^*. (This assumption is not necessary—essentially the same results can be obtained without this assumption—but it avoids some distracting complications.) Note that the usual α-β-γ scheduling cost function, given in Equation (1.1), falls within the class of allowable scheduling cost functions.

We are interested in equilibria: all users departing from $i \in [n]$ should pay a common disutility (considering both their journey time cost and their scheduling cost), and have no options to modify their departure time to incur a lower disutility. We will formalize this later. For now, we make two remarks. First, it is easy to see that mass departures are not possible in an equilibrium (agents leaving just before and just after a mass departure would necessarily incur different disutilities). As such, we do not consider this in our model; a solution can be described by the departure *rate* from each location at each moment in time. Secondly, we can assume that $\tau_{(i,i+1)} = 0$ for all $i \in [n]$, because these values have no impact whatsoever on the equilibrium. Modifying $\tau_{(i,i+1)}$ by some amount only has the effect of changing the actual costs paid by agents departing from $1, 2, \ldots, i$ by the same amount. We make this assumption for the remainder.

6.2.2 The (purely) sensitive demand model

Now consider the following variant. Instead of specifying the total mass Q_i at i, we consider that there are an unbounded mass of agents at each location, whose utility for commuting is some given value C_i. That is, an agent at location i will commute if their total cost for doing so is strictly below C_i, and is indifferent to commuting or not if this cost is exactly C_i, but will not commute at any larger cost.

Define an auxiliary graph G by adding a new source node s, along with arcs (s, i) for $i \in [n]$. Let $C \geq \max_{i \in [n]} C_i$. We give arc (s, i) length $\tau_{(s,i)} := (C - C_i)/\alpha$ and infinite capacity. Observe that a user starting at s at time θ, traversing (s, i), and then departing i at time $\theta + \tau_{(s,i)}$, experiences disutility precisely $\alpha\tau_{(s,i)} = C - C_i$ more than a user departing i at time $\theta + \tau_{(s,i)}$. Thus, an equilibrium solution to the original problem with users at i experiencing disutility C_i maps to a situation in G where all flow experiences disutility C—in other words, an equilibrium of the *single-origin single-destination* instance defined by G. We use the model discussed in Section 4.2, which considers general networks with a single origin and destination, to describe the behavior of equilibria in the augmented network G.

Recall that we assume $\tau_{(i,i+1)} = 0$ for all i (this remains without loss of generality for the sensitive demand model as well, though of course adjusting $\tau_{(i,i+1)}$ requires adjusting C_j for $j = 1, 2 \ldots, i$). As such, we will use τ_i as shorthand for $\tau_{(s,i)}$.

We may assume that $\tau_1 \leq \tau_2 \leq \cdots \leq \tau_n$. The reason is that if $\tau_i > \tau_{i+1}$, the time on the path $(s, i, i+1)$ will always be strictly larger than the direct path $(s, i+1)$ (exploiting that $(s, i+1)$ has infinite capacity and hence no queue). This implies that, in such a case, there will never be flow on (s, i), and node i can be removed from the instance; the arc replacing $(i-1, i)$ and $(i, i+1)$ should be given the smaller capacity of the two arcs.

Notice that adding some constant δ to C and to $\alpha\tau_i$ for all $i \in [n]$ has no impact; the equilibrium remains completely unchanged. We can exploit this symmetry to drop the requirement that free-flow travel times are nonnegative: we can interpret such an instance by shifting values up by a large enough value.

6.3 Equilibria in the sensitive demand model

6.3.1 Equilibrium conditions and existence in the sensitive model

We refer to Section 4.2 for the link dynamics, the earliest arrival functions, the dynamic shortest path networks, the user costs and choices, the equilibrium conditions and the Endogenous Thin Flow with Resetting (ETF).

Observe that users that depart first, as well as users that depart last, experience no queueing delays. This is for the same reason as with a single bottleneck; for first departures, it is of course trivial, and for last departures, if a user encounters a queue, they could deviate to a later starting time without affecting their arrival time, for a strictly smaller disutility. Thus, since the commute time without queues is precisely τ_1, the departure times in an equilibrium are supported on the set

$$[a_1, b_1] := \{\theta : \rho(\theta + \tau_1) + \alpha\tau_1 \leq C\}.$$

We can also determine the arrival time of a user that departs at time $\theta \in [a_1, b_1]$ (if any users do—it will turn out to be the case that users do depart throughout the interval). A user that departs s at time θ and arrives at t at time θ' experiences a total cost of $\alpha(\theta' - \theta) + \rho(\theta')$. Since $\rho'(\theta') > -\alpha$ for all θ', and hence $\xi \to \rho(\xi) + \alpha\xi$ is strictly increasing, we can deduce that θ' is the unique solution of the equation

$$\rho(\theta') + \alpha(\theta' - \theta) = C.$$

Let (f, ℓ) be an equilibrium of the augmented graph G. Notice that the arcs (s, i), for all $i \in [n]$, never grow queues since they have infinite capacity and that all horizontal arcs are active throughout $[a_1, b_1]$. By the ETF conditions (Definition 4.5) and by the specific structure of the network we have that:

$$\ell'_{i+1}(\theta) = \begin{cases} \frac{1}{\nu_i} f_{(i,i+1)}(\theta) & \text{if } q_{(i,i+1)}(\ell_i(\theta)) > 0 \\ \max\{\frac{1}{\nu_i} f_{(i,i+1)}(\theta), \ell'_i(\theta)\} & \text{otherwise} \end{cases} \tag{6.1}$$

for all $i \in [n]$ and almost every θ. Finally, to reach i, either we can go directly from s, taking time τ_i, or we can find the earliest arrival to $i-1$, and then (possibly) spend some time in the queue on arc $(i-1, i)$. This implies that $\ell_i(\theta) = \theta + \tau_i$ if (s, i) is active and that

$$\ell'_i(\theta) = 1 \text{ for almost every } \theta \text{ for which } (s, i) \text{ is active.} \tag{6.2}$$

6.3.2 Equilibrium structure

Throughout this section, we will consider an arbitrary equilibrium f, with associated labels ℓ and queueing delays q.

Define θ^* so that $\ell_t(\theta^*) = T^*$ (recall T^* is the unique minimizer of ρ). We record some simple properties of ℓ_t.

Lemma 6.1. *$\ell'_t(\theta) > 1$ for $\theta < \theta^*$ and $0 < \ell'_t(\theta) < 1$ for $\theta \geq \theta^*$. Further, $\ell'_t(\theta)$ is nonincreasing.*

Proof. This is an immediate consequence of Lemma 4.4, the assumptions on ρ (in particular, that ρ is convex with $\rho'(\theta) > -\alpha$ for almost every θ) and the definition of θ^*. □

We are only interested in the situation where all vertical arcs do become active at some point, and in fact for a strictly positive amount of time; for otherwise, some arcs can simply be removed, without impacting the equilibrium behavior. We call instances that satisfy this property *inclusive*.

Clustering

We can certainly have that some τ_i's are equal, and this will be very important later. However, it is rather easy to see that to understand the equilibrium structure, it will suffice to understand the case of strictly increasing τ_i's, as the following definition and lemmas will make clear.

Lemma 6.2. *For $i > 1$, if $q_{(i-1,i)}(\ell_{i-1}(\theta)) < \tau_i - \tau_{i-1}$, then (s, i) is not active at time θ. In particular, if $\tau_{i-1} < \tau_i$ and there is no queue on $(i-1, i)$ at entrance time θ, then (s, i) is not active at time θ.*

Proof. Under the given conditions,

$$\ell_i(\theta) < (\ell_{i-1}(\theta) - \tau_{i-1}) + \tau_i \leq \theta + \tau_i,$$

where the second inequality is obtained by considering the direct route from s to $i-1$, given that $(s, i-1)$ has no queue. This implies that (s, i) is inactive. □

Lemma 6.3. *All inclusive instances satisfy the following property:*

$$\text{For all } 1 < i < j \leq n \text{ with } \tau_{i-1} < \tau_i = \tau_j,\ \nu_{j-1} \geq \nu_{i-1}. \tag{6.3}$$

Proof. Suppose not; let i, j be a pair violating the property, with j chosen as small as possible. Thus $\nu_{k-1} \geq \nu_{i-1}$ for $i \leq k < j$, since otherwise our choice of j was not minimal. We show that (s, i) is never active.

So fix some time θ. If $q_{(i-1,i)}(\ell_{i-1}(\theta)) = 0$, then (s, i) is inactive at time θ by Lemma 6.2. Otherwise, the outflow from arc $(i-1, i)$ is exactly equal to the capacity ν_{i-1} on an interval $[\ell_i(\theta) - \epsilon, \ell_i(\theta)]$ for some positive ϵ. Since $\nu_{j-1} < \nu_{i-1}$ and $\nu_{k-1} \geq \nu_{i-1}$ for all $i \leq k < j$, the inflow into $j-1$ exceeds ν_{j-1} for some interval $[\ell_{j-1}(\theta) - \epsilon', \ell_{j-1}(\theta)]$, meaning that $q_{(j-1,j)}(\ell_{j-1}(\theta)) > 0$. But then the delay on the path from s to i to j is longer than that of the direct path from s to j, and so (s, i) cannot be active at time θ. So (s, i) is never active, contradicting that the instance is inclusive. □

Lemma 6.4. *If G is inclusive and $\tau_i = \tau_j$, then for the equilibrium labelling ℓ, $\ell_i = \ell_j$.*

Proof. It suffices to show the claim assuming $\tau_{i-1} < \tau_i$ or $i = 1$, by transitivity. Observe that as long as none of the arcs (s, r) for $i \leq r \leq j$ are active, there cannot be a queue on any arc between i and j, since ν_{i-1} is not larger than the capacity of any of these arcs (this holds also when $i = 1$, since then $\nu_0 = 0$). It follows that the first time θ_1 (if any) that any (s, r) for $i \leq r \leq j$ becomes active, all of them become active. However, it is also true that if (s, i) is active at some time θ, then so is (s, j): $\ell_j(\theta) \geq \ell_i(\theta)$, but also $\ell_j(\theta) \leq \theta + \tau_j = \theta + \tau_i = \ell_i(\theta)$. Thus no queues ever form on arcs between i and j, and so $\ell_i(\theta) = \ell_j(\theta)$ for all θ. □

Define the *clustering* of the instance G to be the partition $\mathcal{C}$ of $[n]$ such that $\tau_i = \tau_j$ for all i, j in the same part, and $\tau_i \neq \tau_j$ otherwise. We will call the parts of $\mathcal{C}$ *clusters*. Consider now the instance obtained by contracting each cluster into a single node, which we will denote $G/\mathcal{C}$. By the above lemma, we can unambiguously define the image of ℓ under this contraction, which we will denote $\ell/\mathcal{C}$ (so for any cluster $C \in \mathcal{C}$, $(\ell/\mathcal{C})_C = \ell_i$ for any $i \in C$). We can also define the image $f/\mathcal{C}$ of f under the contraction: if $C = \{i, i+1, \dots, j\} \in \mathcal{C}$, then the flow from s to the cluster C in $G/\mathcal{C}$ at some time θ is given by summing $f_{(s,r)}(\theta)$ for $i \leq r \leq j$.

Lemma 6.5. *If G is inclusive, then $(f/\mathcal{C}, \ell/\mathcal{C})$ is an equilibrium of $G/\mathcal{C}$.*

Proof. By Lemma 6.4, there are no queues in G in the horizontal arcs that are contracted to form $G/\mathcal{C}$. Thus disutilities for any given choice of route and departure time are the same in $(f/\mathcal{C}, \ell/\mathcal{C})$ as in (f, ℓ), and so the equilibrium conditions are satisfied. □

Strictly increasing travel times

We will now discuss the equilibrium structure in the case $\tau_1 < \tau_2 < \cdots < \tau_n$; we will call such instances *strict*. To summarize the main insights about an equilibrium that we will obtain in this section:

- for any $i \in [n]$, the arc (s, i) is active for an *interval* of time;
- the set of times that there is *some* queue to the right of a node i is also an interval, and
- on this interval, the flow $f_{(i,i+1)}$ is almost everywhere nonincreasing, and can be described quite explicitly in a recursive fashion.

Define, for any $\theta \in [a_1, b_1)$, $(\mathsf{r}(\theta), \mathsf{r}(\theta)+1)$ to be the rightmost arc that either has a queue at time θ, or is growing a queue at time θ. This is well-defined; if $\theta \in (a_1, b_1)$, there must be a queue by the equilibrium conditions, and for $\theta = a_1$, some arc must be growing a queue. Also observe that by (6.1),

$$\ell'_{\mathsf{r}(\theta)+1}(\theta) = f_{(\mathsf{r}(\theta),\mathsf{r}(\theta)+1)}(\theta)/\nu_{\mathsf{r}(\theta)} \qquad \text{for almost every } \theta. \tag{6.4}$$

Let

$$\mathcal{R} := \{j \in [n] : \nu_k \geq \nu_j \text{ for all } k \geq j\}. \tag{6.5}$$

Lemma 6.6. *For all $\theta \in [a_1, b_1)$, $\mathsf{r}(\theta) \in \mathcal{R}$.*

Proof. Let $j = \mathsf{r}(\theta)$. Suppose for a contradiction that $k > j$ with $\nu_k < \nu_j$. Let I be an open interval containing θ in which $(j, j+1)$ has a queue throughout. Then by (6.1), $\ell'_j(\theta') = f_{(j,j+1)}(\theta')/\nu_j$ for almost every $\theta' \in I$, whereas $\ell'_k(\theta') \geq f_{(k,k+1)}(\theta')/\nu_k \geq f_{(j,j+1)}(\theta')/\nu_k$. Hence $\ell'_k(\theta') > \ell'_j(\theta')$ for almost every $\theta' \in I$, implying that there is a queue to the right of j throughout I. This contradicts the definition of $\mathsf{r}(\theta)$. □

For $i \in [n]$, let A_i be the set of times where (s, i) is active, and S_i the set of times where there is *some* queue to the right of i:

$$S_i := \{\theta \in [a_1, b_1] : \ell_i(\theta) < \ell_t(\theta)\}$$
$$A_i := \{\theta \in [a_1, b_1] : (s, i) \text{ is active at time } \theta\}.$$

We will eventually show that S_i and A_i are both intervals. Recall that we define θ^* so that $\ell_t(\theta^*) = T^*$.

Lemma 6.7. *A_i is contained in the closure of S_i for each $i \in [n]$.*

Proof. Suppose not; then we must be able to choose a nonempty interval $[\theta_1, \theta_2] \subseteq A_i \setminus S_i$. We can choose θ_1, θ_2 so that either both are at least θ^*, or both are at most θ^*. But then the disutility of a user using (s, i) departing at time θ_1, namely $\rho(\ell_t(\theta_1)) + \tau_i$, differs from that of a user departing at time θ_2, namely $\rho(\ell_t(\theta_2)) + \tau_i$, contradicting the equilibrium conditions. □

Lemma 6.8. *For all $i \in [n]$, S_i is a nonempty open interval whose closure contains θ^*.*

Proof. Since $S_1 = (a_1, b_1)$, the claim holds for $i = 1$. So assume $i > 1$. By Lemma 6.6, if $i - 1 \notin \mathcal{R}$, then $S_{i-1} = S_i$ (the rightmost arc with a queue is never $i - 1$). Thus it suffices to prove the claim assuming that $i - 1 \in \mathcal{R}$.

Let $\theta_1 = \inf S_i$. Then a queue must be growing on some arc $(k, k + 1)$ with $k \geq i$ at time θ_1; since $\nu_j \geq \nu_{i-1}$ for all $j \geq i$, it follows that (s, i) is active at time θ_1 (no arc (s, j) with $j > i$ can be active, since $(j-1, j)$ has no queue at time θ_1.) Thus $\rho(\theta_1+\tau_i)+\alpha\tau_i = 0$. Since $\rho(\theta+\tau_i) < \rho(\theta_1+\tau_i)$ for all $\theta_1 < \theta < \theta^*$, we can deduce from the equilibria conditions that $\ell_i(\theta) < \ell_t(\theta)$ in this same interval, otherwise a strictly improving deviation is possible.

Let $\theta_2 = \inf\{\theta \geq \theta^* : \theta \notin S_i\}$. By the equilibrium conditions, for any $j \geq i$ we have $\rho(\theta_2 + \tau_j) + \alpha\tau_j \geq 0$. Thus, since ρ is strictly increasing from θ^*, $\rho(\theta + \tau_j) + \alpha\tau_j > 0$ for all $\theta > \theta_2$, and hence (s, j) is inactive for all $\theta > \theta_2$. It follows that no queues to the right of i can increase from time θ_2 (using again that $\nu_{i-1} \leq \nu_j$ for all $j \geq i$). Since $\ell_i(\theta_2) = \ell_t(\theta_2)$, we must have that $\ell_i(\theta) = \ell_t(\theta)$ for all $\theta \geq \theta_2$.

We can now deduce that $S_i = (\theta_1, \theta_2)$. Since $[\theta_1, \theta_2]$ contains A_i by Lemma 6.7, which is nonempty by the inclusivity condition, $\theta_1 < \theta_2$. □

We state a useful claim that we will use often. Informally, it tells us that the flow $f_{(i-1,i)}$ on arc $(i - 1, i)$ is strictly above ν_{i-1} just before (s, i) becomes active, is fixed at value ν_{i-1} as long as (s, i) remains active, and then is strictly below ν_{i-1} immediately after (s, i) deactivates.

Claim 6.9. *Suppose (θ_1, θ_2) is an interval in which $(i - 1, i)$ has a queue throughout. Let $\bar{f}$ denote the average value of $f_{(i-1,i)}$ over this interval; that is, $\bar{f} := \frac{1}{\theta_2-\theta_1} \int_{\theta_1}^{\theta_2} f_{(i-1,i)}(\theta)d\theta$. Then:*

- *If (s, i) is active at both times θ_1 and θ_2, $\bar{f} = \nu_{i-1}$.*
- *If (s, i) is active at time θ_2 but not θ_1, $\bar{f} > \nu_{i-1}$.*
- *If (s, i) is active at time θ_1 but not θ_2, $\bar{f} < \nu_{i-1}$.*

Proof. If (s,i) is active at time θ, then $\ell_i(\theta) = \theta + \tau_i$; otherwise, $\ell_i(\theta) < \theta + \tau_i$. So $\ell_i(\theta_2) - \ell_i(\theta_1) - (\theta_2 - \theta_1)$ is zero if (s,i) is active at both θ_1 and θ_2, strictly positive if it is active only at θ_1, and strictly negative if it is active only at θ_2.

Since $(i-1,i)$ has a queue in (θ_1,θ_2), we have that $\ell_i'(\theta) = f_{(i-i,i)}(\theta)/\nu_{i-1}$ for almost every $\theta \in (\theta_1,\theta_2)$, by (6.1). Integrating over the interval and combining with the above observation yields the claim. □

We now come to our main structural lemma. *From now on, we will write (c_i,d_i) for the interval S_i.*

Lemma 6.10. *The following holds for all $i \in [n]$.*

(i) $A_i = [a_i,b_i]$ *for some* $a_i < b_i$.

(ii) $f_{(i,i+1)}(\theta)$ *is an almost everywhere nonincreasing piecewise-constant function on* $S_i = (c_i,d_i)$.

(iii) $a_n < \theta^*$, *and if* $i < n$, $a_i < a_{i+1} \le b_i$.

Proof of Lemma 6.10. We first argue that if (ii) holds for some i, then (i) holds for i as well. Consider two disjoint intervals $(\theta_1-\epsilon,\theta_1)$ and $(\theta_2,\theta_2+\epsilon)$, such that (s,i) is inactive within both intervals, but active at times θ_1 and θ_2. We will show that then necessarily $\theta_1 < \theta_2$; this clearly implies that A_i must be an interval.

The average value of $f_{(i-1,i)}$ over the interval $(\theta_1-\epsilon,\theta_1)$ is strictly larger than over the interval $(\theta_2,\theta_2+\epsilon)$, by Claim 6.9. On these intervals, $f_{(i-1,i)}$ and $f_{(i,i+1)}$ are the same, since (s,i) is inactive and thus sending no flow. By Lemma 6.7, θ_1 and θ_2 are both in the interval $[c_i,d_i]$, and $f_{(i,i+1)}$ is almost everywhere nonincreasing on this interval by assumption. So indeed $\theta_1 < \theta_2$.

We will proceed by induction on i to prove claim (ii). For $i = n$, $f_{(n,n+1)}(\theta) = \ell_t'(\theta)\nu_n$ for all $c_n \le \theta < d_n$; this is nonincreasing by Lemma 6.1. So suppose $i < n$, and that $f_{(i+1,i+2)}$ is nonincreasing on $[c_{i+1},d_{i+1})$. So we also have that $A_{i+1} = [a_{i+1},b_{i+1}]$.

We can easily see that $f_{(i,i+1)}(\theta)$ on the interval $[c_i,d_i)$ can be described as follows, for almost every θ:

$$f_{(i,i+1)}(\theta) = \begin{cases} \ell_t'(\theta)\min\{\nu_i,\dots,\nu_n\} & c_i \le \theta < a_{i+1} \\ \nu_i & a_{i+1} \le \theta < b_{i+1} \\ f_{(i+1,i+2)}(\theta) & b_{i+1} \le \theta < d_{i+1} \\ \ell_t'(\theta)\nu_i & d_{i+1} \le \theta < d_i \end{cases}. \tag{6.6}$$

Here, we are using that

- for all $\theta < a_{i+1}$, by (iii) no arc (s,j) with $j \ge i+1$ is active, and so $\mathsf{r}(\theta)$ must have minimum capacity amongst arcs to the right of i;
- for all $a_{i+1} \le \theta < b_{i+1}$, $(i,i+1)$ has a queue and $\ell_{i+1}'(\theta) = 1$, yielding a flow of ν_i on $(i,i+1)$;
- for all $b_{i+1} \le \theta < d_{i+1}$, $f_{(i,i+1)}(\theta) = f_{(i+1,i+2)}(\theta)$; and

- for all $d_{i+1} \leq \theta < d_i$, $\mathsf{r}(\theta) = i$.

It is clear (also using the inductive assumption) that within each of the subintervals defining (6.6), $f_{(i,i+1)}$ is almost everywhere nonincreasing and piecewise constant. It remains to confirm that the flow is nonincreasing between subintervals, at the moments a_{i+1}, b_{i+1} and d_{i+1} (note that some of these subintervals may be empty).

At times a_{i+1} and b_{i+1}, this is immediate from Claim 6.9, along with the fact that $f_{(i,i+1)}$ is almost everywhere piecewise constant. At time d_{i+1}, there is nothing to prove if either $d_{i+1} = b_{i+1}$ or $d_{i+1} = d_i$, so assume $b_{i+1} < d_{i+1} < d_i$. Then $\mathsf{r}(d_{i+1}) = i$, and so we must have that $i \in \mathcal{R}$ by Lemma 6.6. Now choose $\theta' \in (b_{i+1}, d_{i+1})$ so that the following properties hold:

- $f_{(i-i,i)}(\theta) = f_{(i-1,i)}(\theta')$ for almost every $\theta \in [\theta', d_{i+1})$. This is possible, since $f_{(i-1,i)}$ is piecewise constant almost everywhere in $[c_i, d_i)$.
- None of the arcs $(s, i+1), (s, i+2), \ldots, (s, n)$ are active in the interval $[\theta', d_{i+1})$. Certainly $(s, i+1)$ is not active for $\theta' > b_{i+1}$; for the remaining arcs, observe that Lemma 6.2 implies that (s, j) is inactive if the queue on $(j-1, j)$ is sufficiently small, and there are no queues on any arcs to the right of $i+1$ at time d_{i+1}.

Let $j := \mathsf{r}(\theta')$; $j \geq i+1$ by definition of d_{i+1}. Since $(s, i+1)$ is inactive at time θ', we must have $f_{(i,i+1)}(\theta') = f_{(j,j+1)}(\theta')$. But $f_{(j,j+1)}(\theta') = \ell_t'(\theta')\nu_j$ (by (6.4), since $j = \mathsf{r}(d_{i+1})$). Thus

$$f_{(i,i+1)}(\theta') = \ell_t'(\theta')\nu_j \geq \ell_t'(\theta')\nu_i = f_{(i,i+1)}(d_{i+1}),$$

with the inequality coming from the fact that $i \in \mathcal{R}$ and ℓ_t' is nonincreasing. So $f_{(i,i+1)}$ is indeed nonincreasing almost everywhere on $[c_i, d_i)$.

Finally, we argue (iii). There are two cases.

- **Case 1:** $\nu_i \geq \nu_{i-1}$. Then no queue can form on $(i, i+1)$ until (s, i) becomes active, by Lemma 6.2. If $i = n$, we immediately see that $a_n < \theta^*$, otherwise (s, n) will not be active for an interval of nonzero length.

  So suppose $i < n$. Then $q_{(i,i+1)}(\ell_i(a_{i+1})) \geq \tau_{i+1} - \tau_i > 0$ by Lemma 6.2. But because $\nu_i \geq \nu_{i-1}$, it is impossible to grow a queue on $(i, i+1)$ without sending flow along (s, i). It follows that $a_{i+1} > a_i$. Since $f_{(i,i+1)}(\theta) = \nu_i$ for almost every $\theta \in (a_{i+1}, b_{i+1})$, it follows from (ii) that $f_{(i,i+1)}(\theta) \geq \nu_i$ for almost every $\theta \in (c_i, b_{i+1})$. Thus by Claim 6.9, (s, i) remains active on the interval $[a_i, b_{i+1}]$, and so $b_i \geq b_{i+1} > a_{i+1}$.

- **Case 2:** $\nu_i < \nu_{i-1}$. If $i = n$, then $f_{(i,i+1)}(\theta) = \ell_t'(\theta)\nu_i < \nu_{i-1}$ for almost every $\theta \geq \theta^*$, and so $b_n \leq \theta^*$. Thus by the inclusivity assumption, $a_n < \theta^*$.

  Now suppose $i < n$. Then for almost every $\theta \in (a_{i+1}, d_{i+1})$, $f_{(i,i+1)}(\theta) \leq \nu_i < \nu_{i-1}$. By Claim 6.9, we can deduce that (s, i) is not active within this interval, and hence that $f_{(i-1,i)}$ is equal to $f_{(i,i+1)}$ on this interval. So $f_{(i-1,i)}(\theta) < \nu_{i-1}$ for almost every $\theta \in (a_{i+1}, d_{i+1})$, and hence by Claim 6.9 $b_i \leq a_{i+1}$. Hence $a_i < b_i < a_{i+1}$.

□

6.3.3 Uniqueness and a bound on the number of phases

The following is a fairly immediate consequence of our structural results.

Lemma 6.11. *Any equilibrium of an instance of size n (strict or not) has at most $3n+\omega$ phases, where ω is the number of breakpoints in the scheduling function.*

Proof. It suffices to consider strict instances, given Lemma 6.5.

By Lemma 6.10, $f_{(i,i+1)}$ is almost everywhere piecewise constant on $[c_i, d_i]$. For any θ outside of the interval $[c_i, d_i]$, (s, i) is not active, and so $f_{(i,i+1)}(\theta) = f_{(j,j+1)}(\theta)$, where $j < i$ is chosen maximally so that (s, j) is active. It follows that $f_{(i,i+1)}$ is almost everywhere piecewise constant throughout the evolution. To bound the number of breakpoints of f, observe that by (6.6), any breakpoint of $f_{(i,i+1)}$ within (c_i, d_i) that is *not* a breakpoint of either (i) ρ, or (ii) of $f_{(i+1,i+2)}$ within the interval (c_{i+1}, d_{i+1}), is among a_{i+1}, b_{i+1} and d_{i+1}. It follows that f has at most $3(n-1)$ breakpoints that are not breakpoints of ρ, and hence $3(n-1)+\kappa$ breakpoints in total. This gives a bound of $3n+\kappa-2$ on the number of phases. □

Given this, we can dispense with some awkward technicalities, and also argue that equilibria are unique. In an equilibrium (f, ℓ), on any phase f is almost everywhere constant, and ℓ is linear. It follows that we can modify f on a set of measure zero so that it is simply constant on each phase. This makes f piecewise constant (here we use that there are a finite number of phases) and right-continuous, and ℓ piecewise linear and right-differentiable. It follows by the arguments shown in Section 4.3 that ℓ is unique.

Theorem 6.12. *In any strict instance, both f and ℓ are unique (up to measure zero changes in the case of f). In any instance (not necessarily strict), ℓ is unique.*

Proof. Suppose that $(f^{(1)}, \ell^{(1)})$ and $(f^{(2)}, \ell^{(2)})$ are both equilibria, with $f^{(i)}$ piecewise linear and right continuous for $i = 1, 2$, and suppose for a contradiction that they are distinct. Let $\theta_1 = \inf\{\theta : f^{(1)}(\theta) \neq f^{(2)}(\theta)\}$. Then $\ell^{(1)}(\theta_1) = \ell^{(2)}(\theta_1)$. Also let $\epsilon > 0$ be chosen so that $[\theta_1, \theta_1 + \epsilon]$ is contained within a single phase of both equilibria; so $f^{(1)}$ and $f^{(2)}$ are both constant on this interval. Then $(f^{(1)}(\theta_1+\epsilon), \ell^{(1)\prime}(\theta_1+\epsilon))$ and $(f^{(2)}(\theta_1+\epsilon), \ell^{(2)\prime}(\theta_1+\epsilon))$ are both solutions to the ETF instance (see Definition 4.5 determined by $\ell^{(1)}(\theta_1)$. However, according to Theorem 4.9, this has a unique solution. This means that $f^{(1)}$ and $f^{(2)}$ agree on $[\theta_1, \theta_1 + \epsilon]$, contradicting the maximality of θ_1.

Uniqueness of ℓ follows immediately from uniqueness of f, and by Lemma 6.5, this applies to non-strict instances as well. □

From this point further, we will always assume that f is piecewise constant and right-continuous.

Activation times

In what follows, we will consider the vertical arc delays τ to be variable, keeping all other parameters of the instance (ν and ρ) fixed. As such, we will parametrize some values such as the activation times a_i by τ as well when needed; e.g., $a_i(\tau)$, $b_i(\tau)$.

The main insight from this section is that it will be more convenient to describe a strict instance by the activation times a_i; given a valid choice of activation times, there will be a unique corresponding choice of τ that we can determine if necessary. The main reason that activation times are more convenient is that it turns out that, as long as a_{i-1} is strictly less than a_i,

- the interval $[a_i, b_i]$ depends only on $a_i, a_{i+1}, \dots, a_n$, and
- the flow $f_{(i,i+1)}$ *on the interval* $[c_i, d_i)$ is a function only of $a_{i+1}, \dots, a_n$ (note that there is no dependence on a_i). A subtlety here is that the interval $[c_i, d_i)$ *does* depend on the entire vector of activation times.

This will be absolutely crucial when we return to the insensitive model, since it will allow us to build up an equilibrium "from right to left" in an inductive fashion.

Lemma 6.13. *Let $i \in [n]$ and $\tilde{a}_i < \tilde{a}_{i+1} < \cdots < \tilde{a}_n < \theta^*$ be given. Then there exists $\tilde{f}_{(i,i+1)} : \mathbb{R} \to \mathbb{R}_+$, with $\tilde{f}_{(i,i+1)}$ depending only on $\tilde{a}_{i+1}, \dots, \tilde{a}_n$, so that the following holds.*

> *For any τ for which $a_j(\tau) = \tilde{a}_j$ for all $j \geq i$ and $a_{i-1}(\tau) < \tilde{a}_i$ (if $i > 1$), and taking (f, ℓ) to be the resulting equilibrium,*
>
> *(i) $\tilde{f}_{(i,i+1)}(\theta) = f_{(i,i+1)}(\theta)$ for all $\theta \in [c_i(\tau), d_i(\tau))$, and*
>
> *(ii) $b_i(\tau) = \tilde{b}_i := \max\{\theta : \rho(\ell_t(\theta)) \leq \rho(\ell_t(\tilde{a}_i))$ and $\tilde{f}_{(i,i+1)}(\theta) \geq \nu_{i-1}\}$.*

Proof. For $i = n$, it is of course clear that we can choose $\tilde{f}_{(n,n+1)}(\theta) = \ell_t'(\theta)\nu_n$. The argument for (ii) for $i = n$ follows the argument for general i, so we postpone this. So suppose that $i < n$ and assume inductively that $\tilde{f}_{(i+1,i+2)}$ has been defined and that the claimed properties hold for $i + 1$.

We make the following definition, with $\tilde{d}_{i+1}$ still to be determined, *but only depending on* $\tilde{a}_{i+1}, \dots, \tilde{a}_n$:

$$\tilde{f}_{(i,i+1)}(\theta) = \begin{cases} \ell_t'(\theta) \min\{\nu_i, \dots, \nu_n\} & \theta < \tilde{a}_{i+1} \\ \nu_i & \tilde{a}_{i+1} \leq \theta < \tilde{b}_{i+1} \\ \tilde{f}_{(i+1,i+2)}(\theta) & \tilde{b}_{i+1} \leq \theta < \tilde{d}_{i+1} \\ \ell_t'(\theta)\nu_i & \tilde{d}_{i+1} \leq \theta \end{cases}. \tag{6.7}$$

Then comparing to (6.6), it is clear that as long as $\tilde{d}_{i+1} = d_{i+1}(\tau)$, then $\tilde{f}_{(i,i+1)}(\theta) = f_{(i,i+1)}(\theta)$ on $[c_i, d_i)$. However, we *cannot* choose such a $\tilde{d}_{i+1}$, because d_{i+1} does depend on the entire vector τ. Instead, we will give a definition of $\tilde{d}_{i+1}$ satisfying the following weaker property:

$$\textit{Either } \tilde{d}_{i+1} = d_{i+1}(\tau), \textit{ or } \tilde{d}_{i+1} \geq d_{i+1}(\tau) = d_i(\tau). \tag{6.8}$$

This suffices: in the latter case, the error in our definition of $\tilde{d}_{i+1}$ is harmless within the interval $[c_i(\tau), d_i(\tau))$, and so the definition of $\tilde{f}_{(i,i+1)}$ remains correct.

We consider two cases.

- **Case 1:** $i \notin \mathcal{R}$. Then $\mathsf{r}(\theta)$ is never equal to i for any θ, by Lemma 6.6, and hence $d_i(\tau) = d_{i+1}(\tau)$. Thus we simply set $\tilde{d}_{i+1} = \infty$ in the definition of $\tilde{f}_{(i,i+1)}$, and (6.8) is certainly satisfied.

- **Case 2:** $i \in \mathcal{R}$. We will set $\tilde{d}_{i+1}$ to the first time $\theta > \tilde{a}_i$ that $\ell_{i+1}(\theta) = \ell_t(\theta)$, *under the assumption that* $\ell_i(\theta) < \ell_{i+1}(\theta)$ *for* $\theta \in [b_i(\tau), d_i(\tau))$. We first see how this can be computed, and then argue that this choice for $\tilde{d}_{i+1}$ satisfies the desired property given in (6.8).

  Observe that because $i \in \mathcal{R}$, we must have that $\ell_{i+1}(\tilde{a}_{i+1}) = \ell_t(\tilde{a}_{i+1})$; there is no way any queue could grow to the right of $i+1$ whilst no arc (s, j) for $j \geq i+1$ is active. This information will allow us to determine the moment that all queues to the right of $i+1$ empty out, as follows.

  Consider a user departing at time $\tilde{b}_{i+1} = b_{i+1}(\tau)$ and using the arc $(s, i+1)$. This user must experience the same cost as a user departing at time $\tilde{a}_{i+1}$ using arc $(s, i+1)$, and so

$$\begin{aligned}\rho(\ell_t(\tilde{b}_{i+1})) + \alpha(\ell_t(\tilde{b}_{i+1}) - \ell_{i+1}(\tilde{b}_{i+1}) + \tau_{i+1}) &= \rho(\ell_t(\tilde{a}_{i+1})) + \alpha(\ell_t(\tilde{a}_{i+1}) - \ell_{i+1}(\tilde{a}_{i+1}) + \tau_{i+1}) \\ &= \rho(\ell_t(\tilde{a}_{i+1})) + \alpha\tau_{i+1}.\end{aligned}$$

  Hence

$$\ell_t(\tilde{b}_{i+1}) - \ell_{i+1}(\tilde{b}_{i+1}) = \tfrac{1}{\alpha}(\rho(\ell_t(\tilde{a}_{i+1})) - \rho(\ell_t(\tilde{b}_{i+1}))).$$

  Now since we are assuming that $(i, i+1)$ is resetting for all times $\theta' \in [\tilde{b}_{i+1}, d_i(\tau))$, we can deduce that

$$\begin{aligned}\ell_t(d_{i+1}(\tau)) &= \ell_{i+1}(d_{i+1}(\tau)) \\ &= \ell_{i+1}(\tilde{b}_{i+1}) + \int_{\tilde{b}_{i+1}}^{d_{i+1}(\tau)} \tfrac{1}{\nu_i} f_{(i,i+1)}(\theta')d\theta' \\ &= \ell_{i+1}(\tilde{b}_{i+1}) + \int_{\tilde{b}_{i+1}}^{d_{i+1}(\tau)} \tfrac{1}{\nu_i} \tilde{f}_{(i,i+1)}(\theta')d\theta'.\end{aligned}$$

  From this, we can determine the correct value of $d_{i+1}(\tau)$ under the assumption, and hence fix $\tilde{d}_{i+1}$.

  What if the assumption that does not hold? Since S_i is an interval and $\tilde{a}_i \in S_i$, it must be that $d_i(\tau) = d_{i+1}(\tau)$. Moreover, $\ell'_{i+1}(\theta')$ can only be *larger* than the value of $f_{(i,i+1)}(\theta')/\nu_i$ that we assumed in the above argument, implying that our estimate of $\ell_{i+1}(d_{i+1}(\tau))$ based on the assumption is too small. Thus our estimate of $d_{i+1}(\tau)$ that we used to fix $\tilde{d}_{i+1}$ was too large, and (6.8) is satisfied.

We now come to (ii); here we will allow $i = n$ as well. Let $z = \max\{\theta : \rho(\ell_t(\theta)) \leq \rho(\ell_t(\tilde{a}_i)\}$. If $i-1 \notin \mathcal{R}$, consider $j > i$ minimal with $\nu_{j-1} < \nu_{i-1}$; then $b_i(\tau) \leq a_j(\tau)$, since $f_{(i,i+1)}(a_j(\tau)) \leq f_{(j-1,j)}(a_j(\tau)) = \nu_{j-1}$. This implies, firstly, that $b_i(\tau) \leq \theta^* \leq z$; and secondly, that $\tilde{f}_{(i,i+1)}(\theta) = f_{(i,i+1)}(\theta)$ for $\theta \in [c_i(\tau), b_j(\tau))$. Since $a_j(\tau) < d_i(\tau)$, it follows from Claim 6.9 that indeed $b_i(\tau) = \tilde{b}_i$.

So suppose $i-1 \in \mathcal{R}$. Since $i-1 \in \mathcal{R}$, there are no queues to the right of i before or at time $\tilde{a}_i$. The equilibrium conditions then imply that $b_i(\tau) \leq z$, since a user departing later than z and using arc (s,i) would experience a strictly larger disutility than one departing at time $\tilde{a}_i$. Further, $d_i \geq z$, since otherwise a user departing at time d_i would experience strictly smaller disutility.

Thus $\tilde{f}_{(i,i+1)}$ matches $f_{(i,i+1)}$ on $[c_i(\tau), z)$; since also $b_i(\tau) \leq z$, Claim 6.9 implies that $b_i(\tau) = \tilde{b}_i$. □

Extending this result to non-strict instances is straightforward. Given $\tilde{a}_1 \leq \tilde{a}_2 \leq \cdots \leq \tilde{a}_n$, we have (from Lemma 6.4 and Lemma 6.10 (iii)) that the clustering $\mathcal{C}$ is precisely the partition into equal activation times. Considering the contracted instance, we deduce the following.

Corollary 6.14. *Let $i \in [n]$ and $\tilde{a}_i \leq \tilde{a}_{i+1} \leq \cdots \leq \tilde{a}_n < \theta^*$ be given. Let i' be maximal with $\tilde{a}_i = \tilde{a}_{i'}$. Then there exists $\tilde{f}_{(i',i'+1)} : \mathbb{R} \to \mathbb{R}_+$, with $\tilde{f}_{(i',i'+1)}$ depending only on $\tilde{a}_{i'+1}, \ldots, \tilde{a}_n$, so that the following holds.*

> *For any τ for which $a_j(\tau) = \tilde{a}_j$ for all $j \geq i$ and $a_{i-1}(\tau) < \tilde{a}_i$ (if $i > 1$), and taking (f, ℓ) to be any equilibrium to the instance given by τ,*
>
> *(i) $\tilde{f}_{(i',i'+1)}(\theta) = f_{(i',i'+1)}(\theta)$ for all $\theta \in [c_i(\tau), d_i(\tau))$, and*
>
> *(ii) $b_i(\tau) = b_{i+1}(\tau) = \cdots = b_{i'}(\tau) = \tilde{b}_i$, where $\tilde{b}_i = \max\{\theta : \rho(\ell_t(\theta)) \leq \rho(\ell_t(\tilde{a}_i))$ and $\tilde{f}_{(i',i'+1)}(\theta) \geq \nu_{i-1}\}$.*

We can now observe that the activation vector a uniquely determines τ; as such, we can choose to describe an instance via activation times rather than the vector τ, which we will frequently do in the next section.

Lemma 6.15. *The map $\tau \to a(\tau)$ is 1-1.*

Proof. Since, as already noted, the clustering is fully determined by the activation vector a, it suffices to prove the claim for strict instances. Suppose for a contradiction that $a(\tau) = a(\tau')$, with $\tau \neq \tau'$ being strict and both inducing inclusive instances. Write $a = a(\tau)$ and $a' = a(\tau')$, and let $\mathcal{I}$, $\mathcal{I}'$ refer to the instances associated with τ and τ' respectively.

Let j be minimal such that $\tau_j \neq \tau'_j$; assume without loss of generality that $\tau_j < \tau'_j$. Consider any $\theta < a_j$. None of the arcs (s,i) for $i \geq j$ are active in $\mathcal{I}$ at time θ, by Lemma 6.10 (iii). Since (s,j) is not active in $\mathcal{I}$ and $\tau'_j > \tau_j$, (s,j) is not active in $\mathcal{I}'$, and so again by Lemma 6.10 (i), (s,i) is not active in $\mathcal{I}'$ at time θ. Thus the equilibrium labels are identical for $\mathcal{I}$ and $\mathcal{I}'$ until time a_j. At this point, (s,j) is active in $\mathcal{I}$, but not in $\mathcal{I}'$, and so $a'_j > a_j$.

□

6.4 Existence and uniqueness with insensitive demand

We now turn to the insensitive demand model. For any choice of activation times, we can obtain the equilibrium for the corresponding sensitive demand instance. If it happens to be the case that we choose the activation times precisely right, such that the resulting equilibrium sends precisely the desired mass Q_i from location i for each $i \in [n]$, then we have found an equilibrium for the insensitive instance. The challenge is to determine the correct choice of activation times. In order to show existence for the insensitive model, we must show that a good choice of activation times always exist; and to show uniqueness, that this choice is unique.

Our approach will provide a quite explicit description of the mapping between activation times and demand vectors. More precisely, for a given vector a of activation times, we will find a description for the *set* of demand vectors which can be obtained in an equilibrium for a. The reason that there is in general a set of matching demand vectors, rather than just one, is because we do *not* require a to define a strict instance. This means that equilibria are unique with respect to labels, but not flows; indeed while the total amount of flow departing a single cluster is unique, there is a lot of flexibility in the way in which this flow is distributed between the locations in a cluster. Understandings this structure is one of the main technical challenges.

Once we have this description, we will be able to show that the demand sets corresponding to two different activation vectors are disjoint, which will imply uniqueness. We will also provide an algorithm to find the correct activation time vector for a given demand vector, demonstrating existence at the same time.

Let

$$\mathcal{A} = \{a(\tau) : \tau \text{ defines an inclusive instance}\}.$$

(Note that we do not require τ to be strict). For any $a \in \mathcal{A}$, let $\mathcal{Q}_a$ be the set of all possible demand vectors $(Q_1, \dots, Q_n) \in \mathbb{R}^n_{++}$ that can be obtained in an equilibrium for the instance defined by the activation vector a. (As with our restriction to inclusive instances, the restriction to strictly positive demands is mostly for convenience. If $Q_i = 0$ for some i, we can simply remove i from the instance, and contract whichever of $(i-1, i)$ and $(i, i+1)$ has smaller capacity, to obtain a completely equivalent instance.) We begin by giving a completely explicit description of this set.

Definition 6.16. *For all $0 \le j < r \le n$ and $a_r < a_{r+1} \le a_{r+2} \le \cdots \le a_n$ (note the first inequality uniquely is required to be strict), define*

$$M_{j,r} := \int_{a_r}^{z_r} [\tilde{f}_{(r,r+1)}(\theta) - \nu_j]^+ d\theta. \tag{6.9}$$

Here, $\tilde{f}_{(r,r+1)}$ is as in Corollary 6.14 for the activation times $a_{r+1}, \dots, a_n$, and

$$z_r := \max\Big\{\theta : \rho(\ell_t(\theta)) \le \rho(\ell_t(a_r))\Big\}. \tag{6.10}$$

Note that $M_{j,r}$ depends only on $a_r, a_{r+1}, \dots, a_n$; we will write $M_{j,r}(a_r, a_{r+1}, \dots, a_n)$ if we wish to make this dependence explicit. The interpretation of $M_{j,r}(a_r, \dots, a_n)$ (as will become clear) is the total amount of demand sent from location $j+1, \dots, r$ in a solution

where $j+1, j+2, \dots, r$ are all in the same cluster, but j is not (i.e., $a_{j+1} = a_{j+2} = \cdots = a_r$ and $a_j < a_{j+1}$).

Define, for convenience, $a_{n+1} := \infty$ and $a_0 := -\infty$.

Proposition 6.17. *For any $a \in \mathcal{A}$,*

$$\mathcal{Q}_a = \Big\{ Q \in \mathbb{R}^n_{++} : \sum_{i=j}^{r} Q_i \geq M_{j-1,r} \qquad \forall j \leq r \in [n] : a_j = a_r < a_{r+1},$$
$$\sum_{i=j}^{r} Q_i = M_{j-1,r} \qquad \forall j \leq r \in [n] : a_{j-1} < a_j = a_r < a_{r+1} \Big\}.$$

Further, given any $Q \in \mathcal{Q}_a$, there is an efficient algorithm to find an equilibrium (f, ℓ) corresponding to Q.

Note that as well as specifying precisely how much demand emanates from each cluster (the equality constraints), there are also constraints on how this demand can be distributed within a cluster. Given a cluster these constraints are lower bounds on how much of the demand can come from any rightmost portion of the cluster $\{j, j+1, \dots, r\}$, or equivalently, an upper bound on the demand that can come from any leftmost portion of the cluster. This makes sense intuitively; only during the period when the cluster is active can demand be sent from it, and since the horizontal arcs within the cluster cannot grow queues, they act as bottlenecks to the demand in the portion of the cluster to their left.

Proof of Proposition 6.17. Let $\mathcal{D}$ denote the set of demand vectors satisfying the claimed description of $\mathcal{Q}_a$.

$\mathcal{Q}_a \subseteq \mathcal{D}$. Fix any equilibrium (f, ℓ) of the instance corresponding to a. Consider any $j \leq r \in [n]$ with $a_j = a_r < a_{r+1}$. By Lemma 6.5, we have $b_i = b_r$ for $j \leq i \leq r$, since $\tau_j = \cdots = \tau_r$. The arcs (s, i) for $j \leq i \leq r$ are thus active precisely on the interval $[a_r, b_r]$. We have $f_{(r,r+1)}(\theta) = \tilde{f}_{(r,r+1)}(\theta)$ and $f_{(j-1,j)}(\theta) \leq \nu_{j-1}$ for all $\theta \in [a_r, b_r)$. Moreover, if $a_{j-1} < a_j$, then $f_{(j-1,j)}(\theta) = \nu_{j-1}$ for all $\theta \in [a_r, b_r)$. Hence

$$\begin{aligned}
\sum_{i=j}^{r} Q_i &= \int_{a_r}^{b_r} f_{(r,r+1)}(\theta) - f_{(j-1,j)}(\theta) d\theta \\
&\geq \int_{a_r}^{b_r} \tilde{f}_{(r,r+1)}(\theta) - \nu_{j-1} d\theta \\
&= \int_{a_r}^{z_r} [\tilde{f}_{(r,r+1)}(\theta) - \nu_{j-1}]^+ d\theta \qquad \text{by (6.10)},
\end{aligned}$$

with the inequality being an equality if $a_{j-1} < a_j$. Thus Q satisfies all constraints for $\mathcal{D}$.

$\mathcal{Q}_a \supseteq \mathcal{D}$. Given $Q \in \mathcal{D}$, we wish to demonstrate an equilibrium (f, ℓ) corresponding to a for which Q is the resulting demand vector. We will also make sure to provide an efficient algorithm to construct f. (Once f is known, ℓ is easily computed.)

We will consider each cluster of equal activation times separately. Fix $j, r \in [n]$ with $a_{j-1} < a_j = a_r < a_{r+1}$. We need to determine $f_{(s,i)}(\theta)$ for $j \leq i \leq r$ and $\theta \in [a_r, b_r)$;

this uniquely determines the entire flow via flow conservation. Note that $f_{(j-1,j)}(\theta) = \nu_{j-1}$ and $f_{(r,r+1)}(\theta) = \tilde{f}_{(r,r+1)}(\theta)$ are determined for $\theta \in [a_r, b_r)$. We will work from left to right within this block. Begin with $i = j$. We set $f_{(s,i)}(\theta) = \tilde{f}_{(r,r+1)}(\theta) - f_{(i-1,i)}(\theta)$ for $\theta \in [a_r, \delta_i)$ and $f_{(s,i)}(\theta) = 0$ for $\theta \in [\delta_i, b_r)$, where δ_i is chosen so that $\int_{a_r}^{\delta_i} \tilde{f}_{(r,r+1)}(\theta) - f_{(i-1,i)}(\theta) d\theta = Q_i$. The constraint

$$\sum_{i'=j}^{i-1} Q_{i'} = \sum_{i'=j}^{r} Q_{i'} - \sum_{i'=i}^{r} Q_i \leq M_{j-1,r} - M_{i-1,r}$$

ensures that δ_i exists and is bounded by b_r. Once this has been determined, $f_{(i,i+1)} = f_{(s,i)} + f_{(i-1,i)}$ is determined as well. We increase i and repeat this process, until all flows in the block have been determined.

The resulting flow satisfies flow conservation by construction, and is clearly nonnegative. For all $j \leq i \leq k$ and $\theta \in [a_r, b_r)$, we have $f_{(i,i+1)}(\theta) \leq \nu_{j-1} \leq \nu_i$ (the final inequality by Lemma 6.5). This ensures that no arcs inside the cluster grow a queue: By (6.1), $\ell'_{j+1}(\theta) = \max\{\ell'_j(\theta), f_{(j,j+1)}(\theta)/\nu_j\} = \ell'_j(\theta)$, and similarly for all other nodes in the cluster. □

The following technical "triangle inequality" lemma will be useful.

Lemma 6.18. *For any $a \in \mathcal{A}$ and any $j < k < r$ with $a_k < a_{k+1}$ and $a_r < a_{r+1}$,*

$$M_{j,k} + M_{k,r} \geq M_{j,r}.$$

Proof. It suffices to prove the claim under the assumption that $a_{k+1} = a_r$, since repeated applications then yield the claim in its full generality. So we assume $a_{k+1} = \cdots = a_r$ and hence (by Lemma 6.5) $b_{k+1} = \cdots = b_r$ and $\nu_k \leq \nu_{r-1}$. We have

$$\begin{aligned} M_{j,k} + M_{k,r} &= \int_{a_k}^{z_k} [\tilde{f}_{(k,k+1)}(\theta) - \nu_j]^+ d\theta + \int_{a_r}^{z_r} [\tilde{f}_{(r,r+1)}(\theta) - \nu_k]^+ d\theta \\ &\geq \int_{a_r}^{z_r} [\tilde{f}_{(k,k+1)}(\theta) - \nu_j]^+ d\theta + [\tilde{f}_{(r,r+1)}(\theta) - \nu_k]^+ d\theta \qquad \text{(since } a_k < a_r \text{ and } z_k \leq z_r\text{)} \\ &\geq \int_{a_r}^{z_r} [\tilde{f}_{(r,r+1)}(\theta) - \nu_j]^+ d\theta = M_{j,r}. \end{aligned}$$

Here, the last inequality follows by observing that:

- either $\theta < b_r$, in which case $\tilde{f}_{(k,k+1)}(\theta) = \nu_k$, and the triangle inequality for $[\cdot]^+$ can be applied,
- or $\theta \geq b_r$, in which case $\tilde{f}_{(k,k+1)}(\theta) = \tilde{f}_{(r,r+1)}(\theta)$ and the first term alone suffices.

□

Next, we observe some additional constraints satisfied by demand vectors in $\mathcal{Q}_a$. Note that in the following, no restrictions on j aside from $j \leq r$ are present.

Lemma 6.19. *For any $a \in \mathcal{A}$, $Q \in \mathcal{Q}_a$, and $j \leq r \in [n]$ with $a_r < a_{r+1}$,*

$$\sum_{i=j}^{r} Q_i \geq M_{j-1,r}.$$

Proof. We proceed by induction on j; the claim clearly holds whenever $a_j = a_r$, and in particular when $j = r$. Otherwise, let $k \leq r$ be minimal such that $a_k > a_j$. Then inductively, $\sum_{i=k}^{r} Q_i \geq M_{k-1,r}$. We also have $\sum_{i=j}^{k-1} Q_i \geq M_{j-1,k-1}$ by Proposition 6.17. The claim then follows from Lemma 6.18. □

We now describe an algorithm that, given $Q \in \mathbb{R}^n_{++}$, determines an activation time vector $a \in \mathcal{A}$ such that $Q \in \mathcal{Q}_a$. We will show that this algorithm always succeeds, and also show that the returned vector a is the unique possible choice.

Algorithm 1 Determining an activation time vector a with $Q \in \mathcal{Q}_a$.

Require: Demand vector $Q \in \mathbb{R}^n_{++}$.
Ensure: Activation vector $a \in \mathcal{A}$ so that $Q \in \mathcal{Q}_a$.

$r \leftarrow n$.
while $r > 0$ **do**
  Choose y minimally such that $\sum_{i=j}^{r} Q_i \geq M_{j-1,r}(y, a_{r+1}, \ldots, a_n)$ for all $j \leq r$.
  Choose k minimally such that $\sum_{i=k}^{r} Q_i = M_{k-1,r}(y, a_{r+1}, \ldots, a_n)$.
  Set $a_k = a_{k+1} = \cdots = a_r = y$.
  $r \leftarrow k - 1$.
end while
return a.

Theorem 6.20. *Algorithm 1 always succeeds, for any $Q \in \mathbb{R}^n_{++}$.*

Proof. We will proceed by induction. We will show that the following invariant holds throughout the algorithm, whenever $r < n$:

$$\sum_{i=j}^{r} Q_i > M_{j-1,r}(a_{r+1}, a_{r+2}, \ldots, a_n) \qquad \forall j \leq r. \tag{6.11}$$

So assume that $a_{r+1}, \ldots, a_n$ have been fixed. Note that $M_{j-1,r}(y, a_{r+1}, \ldots, a_n)$ is a continuous and decreasing function of y. Given that the invariant holds for r (or $r = n$), we can deduce that $y < a_{r+1}$, and also the existence of the index k (if no constraint is tight, y cannot be minimal). By the choice of k, $\sum_{i=k}^{r} Q_i = M_{k-1,r}(a_r, \ldots, a_n)$ and $\sum_{i=j}^{r} Q_i \geq M_{j-1,r}$ for all $k < j \leq r$. Thus (taking into account the induction) all constraints of the description of $\mathcal{Q}_a$ indexed by $j' \leq r'$ with $r' \geq r$ are satisfied.

Finally, we observe that the invariant (6.11) holds for the next iteration. Applying Lemma 6.18, we deduce

$$\sum_{i=j}^{k-1} Q_i = \sum_{i=j}^{r} Q_i - \sum_{i=k}^{r} Q_i > M_{j-1,r} - M_{k-1,r} \geq M_{j-1,k},$$

as required. □

Theorem 6.21. *If $a \neq a' \in \mathcal{A}$, then $\mathcal{Q}_a \cap \mathcal{Q}_{a'} = \emptyset$.*

Proof. Suppose for a contradiction that $Q \in \mathcal{Q}_a$ and $Q \in \mathcal{Q}_{a'}$. We'll use $M_{j,r}$ as shorthand for $M_{j,r}(a_r, \ldots, a_n)$ and $M'_{j,r}$ as shorthand for $M_{j,r}(a'_r, \ldots, a'_n)$.

Let k be chosen maximally so that $a_k \neq a'_k$; assume, without loss of generality, that $a'_k < a_k$. Choose r so that $a_k = a_r < a_{r+1}$, and j so that $a_{j-1} < a_j = a_k$.

First, we observe that $M'_{j-1,k} > M_{j-1,k}$. Since $Q \in \mathbb{R}^n_{++}$, $M'_{j-1,k} > 0$; exploiting that $\tilde{f}_{(k,k+1)}(\theta)$ is decreasing, we must then have that $\tilde{f}_{(k,k+1)}(a_k) > \nu_{k-1}$. By Lemma 6.5, $\nu_{j-1} \leq \nu_{k-1}$. Hence $\int_{a'_k}^{a_k} [\tilde{f}_{(k,k+1)}(\theta) - \nu_{j-1}]^+ d\theta > 0$, and so $M'_{j-1,k} > M_{j-1,k}$.

By Lemma 6.19 applied to $\mathcal{Q}_{a'}$, $\sum_{i=j}^{k} Q_i \geq M'_{j-1,k} > M_{j-1,k}$. Finally, applying Lemma 6.18,

$$\sum_{i=j}^{r} Q_i = \sum_{i=j}^{k} Q_i + \sum_{i=k+1}^{r} Q_i > M_{j-1,k} + M_{k,r} \geq M_{j-1,r},$$

contradicting (via Proposition 6.17) the assumption that $Q \in \mathcal{Q}_a$. □

6.4.1 Algorithmic issues

We now discuss the details of implementing Algorithm 1 and computing an actual equilibrium flow, and the running time of the algorithm.

Theorem 6.22. *The time required by Algorithm 1 is $O(n(n+\omega))$. Furthermore, all the values (f, ℓ) can be computed in $O(n(n+\omega))$ time.*

Proof. For the first part of the statement, our goal is to show that Algorithm 1 runs at most n iterations, each with amortized complexity $O(n+\omega)$.

For each choice of r we compute the piecewise constant function $\tilde{f}_{r,r+1}$. For this we proceed as in the proof of Lemma 6.13, specifically as in Equation (6.7), by setting $\tilde{d}_{r+1}$ to infinity if $r \notin \mathcal{R}$ or, otherwise, to the first time $\theta > \tilde{a}_i$ that $\ell_{i+1}(\theta) = \ell_t(\theta)$, under the assumption that $\ell_i(\theta) < \ell_{i+1}(\theta)$ for $\theta \in [b_i(\tau), d_i(\tau))$. Since the number of phases is at most $3n+\omega$ (Lemma 6.11), it follows that $\tilde{f}_{(r,r+1)}$ can assume at most $3n+\omega$ values. This, together we the fact that $\tilde{f}_{(r+1,r+2)}$ was previously computed, implies that $\tilde{f}_{(r,r+1)}$ requires $O(n+\omega)$ time (for the case $r = n$ we just need to examine ρ, requiring thus only $O(\omega)$ time).

Once we have $\tilde{f}_{r,r+1}$, we can compute the piecewise linear function $\tilde{F}_{r,r+1}(\theta) := \int_{a_1}^{\theta} \tilde{f}_{r,r+1}(\theta')d\theta'$ in time $O(n+\omega)$.

To compute $M_{j-1,r}$, for $j \leq r$, in addition to $\tilde{F}_{r,r+1}(\theta)$ we also need to calculate z_r. Let $\varsigma : \mathbb{R} \to \mathbb{R}$ be the function for which $\varsigma(a_r) = z_r$. This function is also piecewise linear and to compute it we need to examine ρ, requiring thus $O(\omega)$ time. Additionally, it does not depend on r and thus can be computed just once.

Let z'_j be the point at which $\tilde{f}_{r,r+1}(\theta)$ drops below ν_{j-1}, and let $\varsigma_j(y) = \min\{\varsigma(y), z'_j\}$. Notice that

$$M_{j-1,r}(y, a_{r+1}, \ldots, a_n) = \tilde{F}_{r,r+1}(\varsigma(y)) - \tilde{F}_{r,r+1}(y) - \nu_{j-1} \cdot (\varsigma(y) - y). \tag{6.12}$$

Now we show how to determine the values of y and k satisfying the properties indicated in Line 3 and 4 of Algorithm 1. We start with $k = r$ and we proceed as follows:

- We compute the value of y such that $Q_r = M_{r-1,r}(y, a_{r+1}, \dots, a_n)$. This also requires $O(n+\omega)$ time: from Equation (6.12), from the piecewise linearity of ς and from the fact that $\tilde{F}_{r,r+1}(\theta)$ has at most $3n+\omega$, it follows that $y \to M_{k-1,r}(y, a_{r+1}, \dots, a_n)$ is piecewise linear with $O(n+\omega)$ breakpoints. We compute these breakpoints and then examine them in order, until we find the interval containing the solution we seek. At that point we perform a linear interpolation to find the exact value of y.

- Once we have this y, we check if the constraints $\sum_{i=j}^{r} Q_i \geq M_{j-1,r}(y, a_{r+1}, \dots, a_n)$ hold for all $j < k$ (and also, whether they are tight). This can be done in constant time using Equation (6.12).

- If the constraints do hold, then we know that we have found the correct value of y, and we set k minimally so that the constraint is tight. If not, then we know that $k \leq r-1$. Therefore, we set $k = r-1$ and we repeat the procedure.

This means that the total work we do for any fixed value of k is $O(n+\omega)$, and since there are only n values of k, we get a complexity of $O(n(n+\omega))$ for Algorithm 1.

For the second part of the statement, once we have the activation vector, the entire equilibrium (f, ℓ), as well as the corresponding τ vector, can be easily computed in time $O(n(n+\omega))$. First, we can compute the flows $f_{(i,i+1)}$ in their entirety for all i with $a_i < a_{i+1}$ (and $i = n$): for $\theta \in [c_i, d_i)$, $f_{(i,i+1)}(\theta) = \tilde{f}_{(i,i+1)}(\theta)$, with $\tilde{f}$ already computed during Algorithm 1; and for θ not in this interval, $f_{(i,i+1)}(\theta) = f_{(j,j+1)}(\theta) = \tilde{f}_{(j,j+1)}(\theta)$, where j is maximal such that (s, j) is active. For horizontal arcs within a cluster, where $a_i = a_{i+1}$, the proof of Proposition 6.17 provides one possible choice of $f_{(i,i+1)}$.

Once f has been fully determined, it is straightforward to compute ℓ: for each phase, (6.1) provides the derivative of ℓ. The vector τ corresponding to a (and hence the disutility experienced at each location) can then be determined from the equilibrium conditions. □

6.5 Some examples of equilibrium behavior

In this section we present three examples of instances that provide nice insights into the equilibrium behavior. In all the examples we assume the most common scheduling cost function, the one defined in Equation (1.1), with $\alpha = 2$, $\beta = 1$ and $\gamma = 4$ (so $\ell_t'(\theta) = 2$ for $\theta < \theta^*$ and $\ell_t'(\theta) = 1/3$ for $\theta > \theta^*$). The difference among the instances lie in the arc capacities of the network. In one all the horizontal arcs have same capacity, in an other they are decreasing and in the last one they are increasing.

In all the examples we will first discuss the behavior in the case that the clustering is trivial—that is, the sensitive instance that corresponds to the given demand vector is strict. Alternatively, it can be viewed as discussing the behavior on the appropriate contracted instance. As such, each individual location should be understood to potentially represent a number of locations in the original instance. The clustering itself will depend on the demand vector Q. As we have seen, this correspondence is quite complicated, but we will make some qualitative observations, in particular regarding what clusterings are possible in each of the three examples.

All corridor capacities equal. Here we consider the setting in which all horizontal arcs have equal capacity, which we take to be one without loss of generality. Figure 6.1 shows the evolution of an equilibrium in a typical instance of this type, with a trivial clustering. For this restricted setting, the structure is simplest; in addition to the structural results already discussed that hold in general, the following key properties hold.

- Aside from $(s, 1)$, which sends flow throughout, there is *at most* one other vertical arc with positive flow at any moment in time.
- *The first part of the peak.* Let $(s, i(\theta))$ denote the rightmost vertical arc sending flow at time θ, for $\theta \in [a_1, \theta^*)$. Then $i(\theta)$ is an increasing function of θ, and all arcs $(s, 1), (s, 2), \ldots, (s, i(\theta))$ are active (though only $(s, 1)$ and $(s, i(\theta))$ are sending flow).

  The queue on the arc $(i(\theta), i(\theta) + 1)$ increases, and all other queues remain constant.
- *The second part of the peak.* After θ^*, the only active vertical arc (and hence only location sending flow) is $(s, 1)$. Queues empty out from left to right: a queue only begins to empty once it is the leftmost remaining queue.

Let us observe how these follow from our previous structural insights. Consider the situation in the first part of the peak, and assume inductively that $(s, 1), (s, 2), \ldots, (s, i)$ are all active, and that none of the horizontal arcs to the right of $i + 1$ have a queue; here $i = i(\theta)$. Then all horizontal arcs to the left of i must have a queue (by Lemma 6.2), and the flow on all these arcs must be 1, the corridor capacity. Thus $f_{(s,1)}(\theta) = 1$ and $f_{(s,j)}(\theta) = 0$ for all $1 < j < i$. All queues to the left of i will thus remain constant. On the other hand, $(i, i + 1)$ will be growing a queue; $\ell'_{i+1}(\theta) = \ell'_t(\theta) > 1$. The current phase will only end due to reaching θ^*, or a new vertical arc becoming active; this can only be $(s, i + 1)$. Thus the claimed structure is preserved in all phases of the first part of the peak.

At time θ^*, all horizontal arcs have queues. But now $\ell'_t(\theta) < 1$ for $\theta < \theta^*$, and $f_{(n,n+1)}(\theta) = \ell'_t(\theta) < 1$. Since then $f_{(i,i+1)}(\theta) < 1$ for all i, we must have that $\ell'_i(\theta) < 1$ for all $i > 1$. So no vertical arcs aside from $(s, 1)$ can remain active after time θ^*, and all flow departs via $(s, 1)$ from this point forward, at rate below 1. It is clear that queues will then empty from left to right; as long as $(i - 1, i)$ has a queue, the outflow from this arc has unit rate, and so the queue on $(i, i + 1)$ will not change.

We now consider possible clusterings. Here, they are very restricted: only the leftmost cluster may be nontrivial; all other clusters must be singletons. For suppose $\{j, j+1, \ldots, r\}$ were a cluster of the sensitive instance corresponding to a demand vector $Q \in \mathbb{R}^n_{++}$, with $1 < j < r$. Because all capacities are equal, we have that $M_{j-1,r} = M_{r-1,r}$. But then by Proposition 6.17, $\sum_{i=j}^{r} Q_i = M_{j-1,r} = M_{r-1,r} \leq Q_r$. This implies that $Q_j = Q_{j+1} = \cdots = Q_{r-1} = 0$, a contradiction.

So perhaps surprisingly, in this case we have a potentially large group of locations, far from the CBD, which experience the same disutility. As we proceed towards the CBD from this cluster, the disutility strictly decreases.

The size of this single nontrivial cluster will of course depend on the demand vector. While there is no clear way to determine it aside from executing the algorithm, one can

get some intuition from the following thought experiment. Suppose we fix $Q_2, \ldots, Q_n$ and vary Q_1. As long as Q_1 is large enough, location 1 will form a singleton cluster (this can be argued from Proposition 6.17). Conversely, keeping in mind that $(s, 1)$ is active throughout the evolution, if Q_1 is very small then this will not be possible; 1 will be part of a larger cluster. Generally, as Q_1 is decreased, the cluster containing 1 will grow monotonically. All of this can formally be deduced from Proposition 6.17.

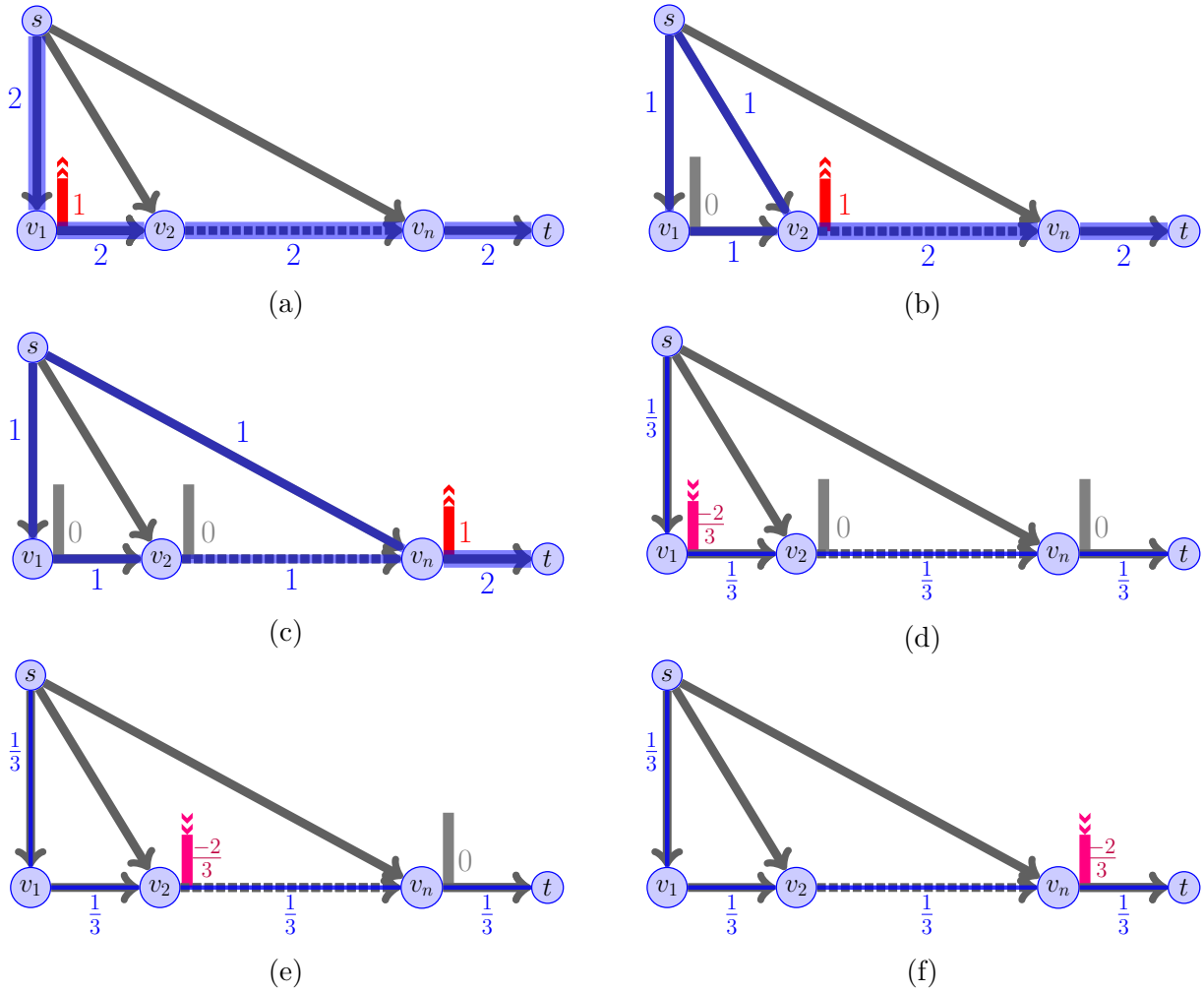


Figure 6.1: Evolution of the equilibrium of an instance with all capacities equal. Times θ are chronological, with a to cbeing for $\theta < \theta^*$, and d to f being for $\theta > \theta^*$. *Legend: Values in blue indicate the flow values $f_e(\theta)$; values in red, purple and gray show the rate of change of queue waiting times. Red is used for increasing queues, purple for decreasing queues, and gray for queues whose length is staying constant.*

Corridor capacities decreasing towards the CBD. We now consider the setting where capacities decrease monotonically from left to right: $\nu_{i+1} < \nu_i$ for all $i \in [n-1]$. As before, we begin the discussion assuming that all clusters are singletons. This implies that in fact the decrease in capacity from left to right is not *too* large: $\nu_n > \nu_1/2$. For otherwise, a queue on $(1, 2)$ at some time θ would imply $\ell_t'(\theta) > 2$, which is not possible.

Figure 6.2 shows a typical equilibrium. Similarly to the setting of equal capacities, it remains true that only $(s, 1)$ as well as at most one other vertical arc are sending flow at any one time. However, we now have that *all* horizontal arcs grow queues at the very start of the evolution. Indeed, as long as some arc $(i-1, i)$ does have a queue, the arc immediately to the right will be growing a queue, since the outflow of arc $(i-1, i)$ is greater than the capacity of $(i, i+1)$. Queues *can* begin to dissipate during the first part of the peak, however; necessarily, this occurs from left to right.

In terms of possible clusterings, the situation is essentially the same as for the equal capacity case: the only possible nontrivial cluster is of the form $\{1, 2, \ldots, j\}$. The reason is essentially the same, exploiting this time that $M_{j-1,r} \leq M_{r-1,r}$ for any $j \leq r$.

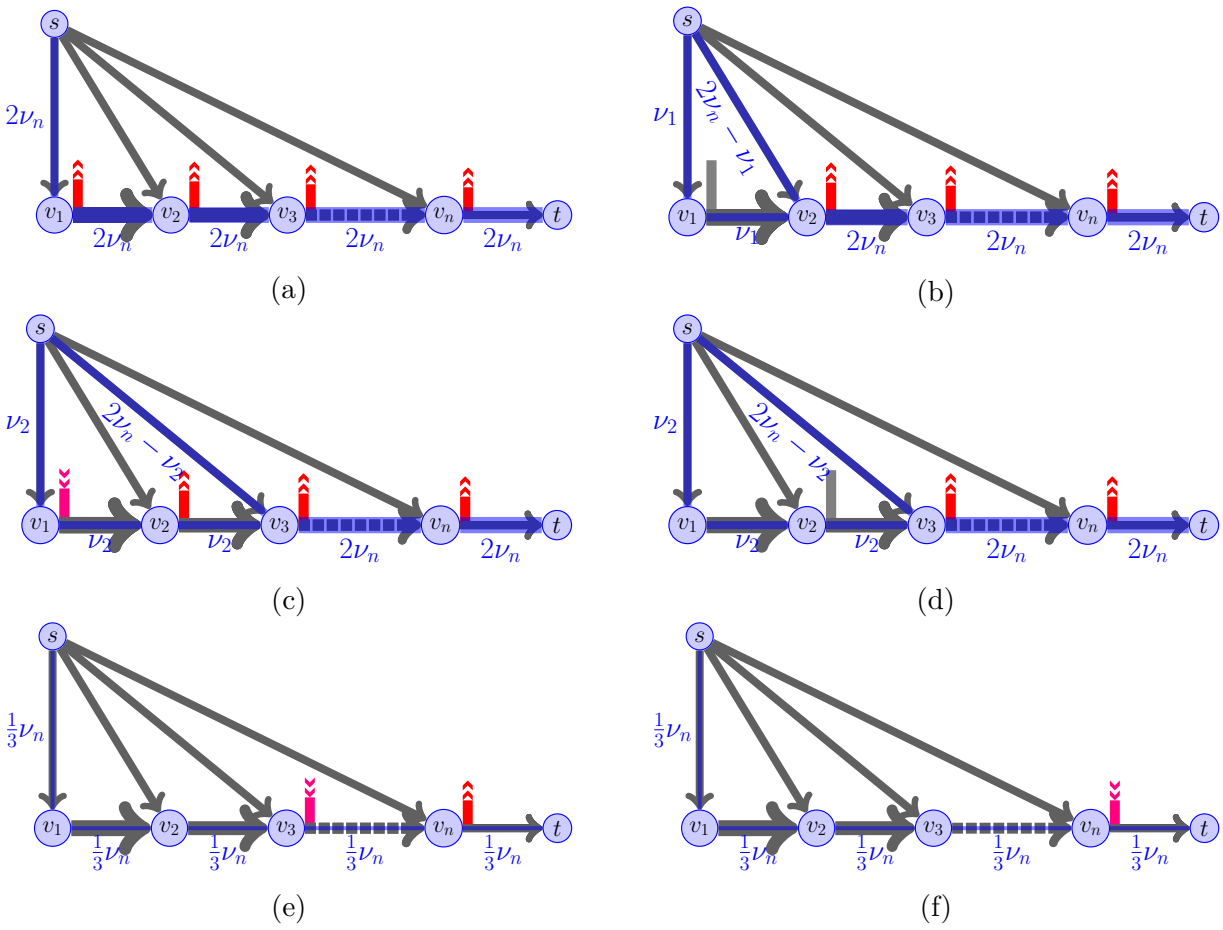


Figure 6.2: Chronological evolution for an instance where capacities decrease towards the CBD. a to d show behavior before θ^*, and e to f behavior after. *Legend: See Figure 6.1.*

Corridor capacities increasing towards the CBD. Now consider the situation where corridor arcs have strictly increasing capacity from left to right. See Figure 6.3 for a typical evolution (again, with all singleton clusters). During the first part of the peak, all but one queue remains constant, and a single queue increases in length. This growing queue moves from left to right as the equilibrium evolves. But now flow is sent

along many vertical arcs: all locations to the left of the growing queue will be sending flow (and none of the locations to the right).

During the second part of the peak, the situation is very different to the previous two settings. We may have flow on vertical arcs aside from $(s, 1)$. If at some moment (s, i) is the rightmost active vertical arc, then all queues to the left of i will remain constant, while queues to the right will dissipate. However, these dissipating queues need not empty in any particular order; this will depend on the specifics of the instance.

This time, there is much more flexibility in the possible clusterings—essentially any clustering is possible, depending on the demand vector Q.

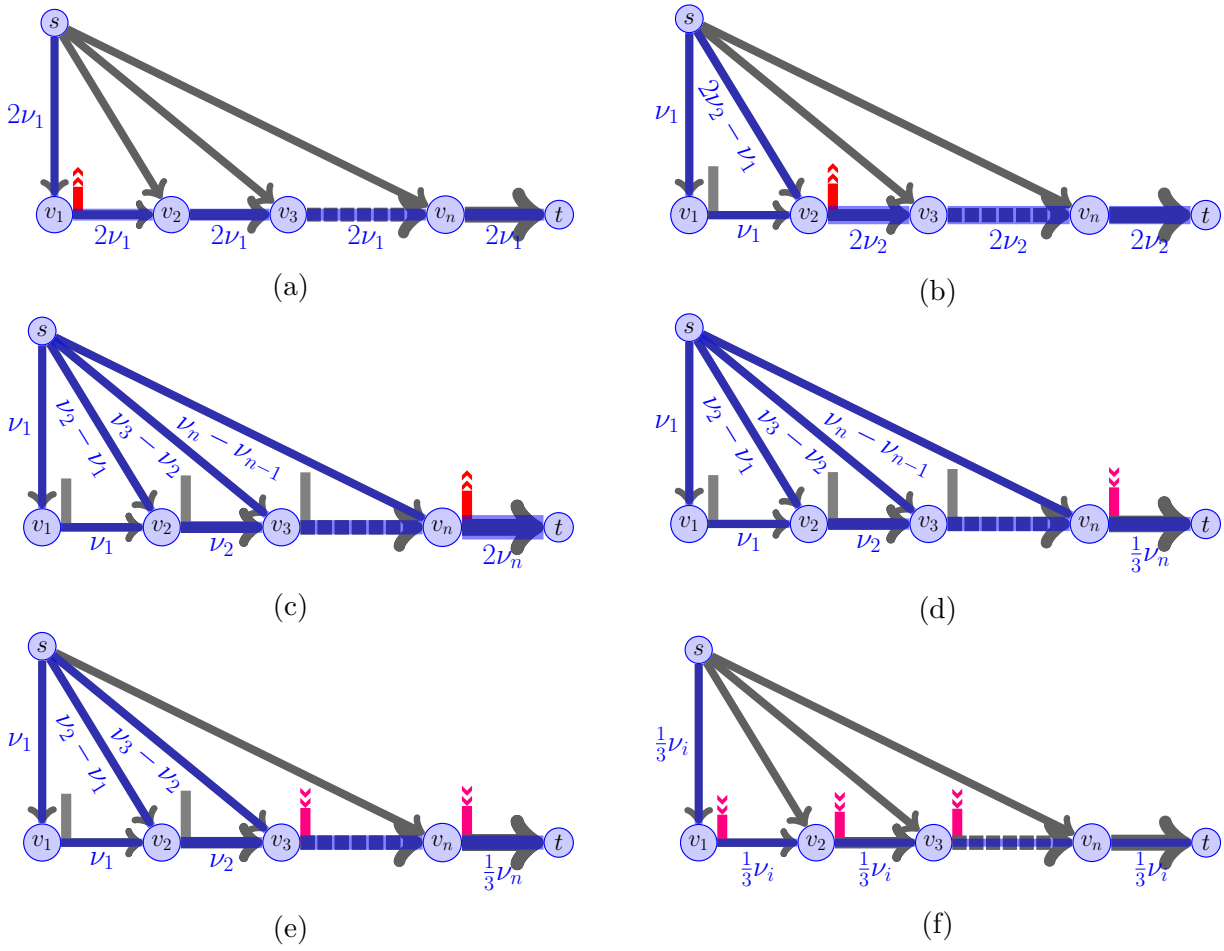


Figure 6.3: Chronological evolution of the equilibrium in an instance with arcs capacities that increase towards the CBD. a to c show behavior before θ^*, and d to f show behavior after. *Legend: See Figure 6.1.*

Comparison with the Arnott-DePalma corridor model. Let us now recall the corridor model in the form introduced by Arnott and DePalma [2011b], where the LWR model is used for modelling traffic congestion, and residential density is distributed continuously in space. They consider only the situation in the first part of the peak, and with

uniform corridor capacity; so we will compare only with the Vickrey corridor model with all corridor capacities equal (we denote their common capacity by ν_1).

Arnott and DePalma make a number of observations about equilibrium properties in their model; here we discuss some of these properties, and what qualitative similarities there are with our structural results. They pay particular attention to the shape of the *departure set*, which is the set of pairs (x, θ), where x is a location in the corridor and θ a moment in time, for which users departing at position x leave at time θ.

- *The upper boundary of the departure set is a vehicle trajectory.* A cursory look back at Figure 6.1 shows that is clearly not the case in the Vickrey corridor model. However, there is a sense in which it is almost true: all vertical arcs are active at time θ^*, the end of the first part of the peak. If we perturb the capacities very slightly, so that they increase every so slightly as one gets closer to the CBD, then we instead have the situation shown in Figure 6.3, where indeed all locations (more precisely, all clusters) are sending flow at time θ^*.

- *At any location, the departure set at that location is an interval. Further, the departure set is connected and does not contain holes.* Consider a trivial clustering. We have seen that this is indeed the case (even more generally for arbitrary corridor capacities).

  With a nontrivial clustering, the flexibility we have in choosing the departure pattern within a cluster means that this is generally not true; it is necessary to consider the contracted instance instead.

- *At any location, the departure rate in the departure set is constant.* (This is slightly modified from the statements in [Arnott and DePalma, 2011b], where there is some difficulties involving the lower boundary that we can ignore for this discussion.) If the clustering is trivial, then this does hold. For consider some (s, i): for the interval in which flow is sent from (s, i), (s, i) is the rightmost active vertical arc. Thus $f_{(s,i)}(\theta) = f_{(i,i+1)}(\theta) - f_{(i-1,i)}(\theta) = \nu_1 \ell_t'(\theta) - \nu_1$ throughout this interval. (This fails the moment corridor capacities are even slightly unequal, however; witness again Figure 6.3.)

  If the clustering is not trivial, then the claim only holds if the aggregate flow from an entire cluster, rather than a single location in the cluster, is considered.

- *Horn-shaped departure set.* Consider the lower boundary of the departure set in the Arnott-DePalma corridor model. They argue that the earliest departure time from a location must be larger for locations closer to the CBD. This holds in the Vickrey corridor model as well.

6.6 Conclusion

In this chapter we used the model and the algorithm discussed in Chapter 4 to study a particular network topology – the one of the corridor model.

We proved existence and uniqueness of the equilibrium, we developed a polynomial time algorithm for computing it and we provided a description of its structure.

Our results showed that, despite the simplicity of the network topology, the equilibrium behaviour is surprisingly involved.

One of the main insight of this work is that the sensitive version of the model is actually easier to deal with and we can completely separate out the additional difficulties needed to map the relationship between the sensitive and insensitive models. An interesting research direction might be exploring this methodology on more general topologies, like trees (multiple origins and single destination) for example.

Finally notice that, following an approach similar to the one of Chapter 5, one could compare the standard (untolled) equilibrium with the one under optimal pricing; the former computed with the algorithm presented in this chapter and the latter with the one of Chapter 3. This could provide interesting insights on the impact of pricing in this network topology and it might be another interesting research direction.

Chapter 7

Long term behavior of dynamic equilibria with exogenous departure time choice

This chapter contains unpublished results obtained thanks to the collaboration of Neil Olver.

7.1 Introduction

In this chapter we study the long term behavior of *dynamic equilibria with exogenous departure time choice* – equilibria where users do not choose their departure time but are released at a constant inflow rate. The cost of a user is just its journey time and, as a consequence, a dynamic equilibrium occurs when each particle travels along a shortest path, taking into account congestion delays caused by other users. This model, as discussed in Section 1.2, was characterized by Koch and Skutella [2011] and further elaborated by Cominetti et al. [2015].

Cominetti et al. [2017] showed that the equilibrium flow reaches a *steady state* in finite time when the inflow rate into the network is a constant not bigger than the capacity of the minimum s-t-cut of the network[1]. They define a steady state to be a nonending time interval in which the queue delay and transit time of any link do not change. With this definition the assumption on the inflow rate into the network is a necessity since, otherwise, there would always be some queue growing.

In this chapter we slightly modify the definition of steady state so that it still matches the one of [Cominetti et al., 2017] when the inflow rate into the network is not bigger than the capacity of the minimum s-t-cut but that can also be applied when the inflow rate is bigger. Namely, we say that a steady state is a nonending time interval in which the rate of change of the transit time of any link does not change. This means that the queue lengths either remain constant or increase linearly over time.

[1]Given a graph where s ant t are connected, i.e. where there is a path from s to t, an s-t-cut C is a subset of the arcs whose removal disconnects s from t. The capacity of a cut is the sum of all the capacity of the arcs of C, namely $\sum_{e \in C} \nu_e$. A minimum s-t cut is an s-t cut with minimum capacity.

With this new definition we conjecture that an equilibrium with constant inflow rate always reaches a steady state.

Our results. In this chapter we drop the assumption on the inflow rate into the network and we provide a more general characterization of steady states which lead to the definition of a potential function that is quite natural and that generalizes the one used by Cominetti et al. [2017]. However, we could not prove that the potential function is always monotone along the evolution of the equilibrium but, instead, we found a network and a queue lengths configuration for which the potential function is not monotone. The queue lengths configuration is quite unnatural: the shortest path network consists in only one s-t-path where all the arcs host a queue, even the arcs that have bigger capacity than their predecessors. We do not know if the equilibrium ever attains such configuration and we suspect that further insights into the relation between the thin flows of consecutive shortest path networks are needed in order to solve this problem.

Outline of the chapter. We formally define the model in Section 7.2. In Section 7.3 we provide the characterization of steady states. Finally in Section 7.4 we introduce the aforementioned potential function and counterexample.

7.2 Model and preliminaries

The link dynamics, the earliest arrival functions and dynamic shortest path network are defined as in Section 4.2.

User choices, dynamic equilibria and induced cumulative flow. As in the endogenous case, each individual user is considered to control a negligible fraction of the total flow, so that there are an infinite number of infinitesimally small, divisible users. Starting from time 0, users are released into s over time at a constant rate ν_0 and each user seeks to arrive at the sink t as early as possible. Thus the cost of a particle is simply its journey time and we have an equilibrium when each particle travels along a path in the shortest path network.

For this setting, Condition 4.1, Equation (4.7) and Lemma 4.2 of Section 4.2 fully characterize the equilibrium flow [Cominetti et al., 2015].

Derivatives of a dynamic equilibrium. For almost every θ, let $x' \coloneqq \frac{dx}{d\theta}$ and $\ell' \coloneqq \frac{d\ell}{d\theta}$ be the derivatives of the cumulative flow $(x_e)_{e\in E}$ and the earliest arrival functions $(\ell_v)_{v\in V}$. By differentiating Equation (4.7) we see that $x'(\theta)$ is a static flow. Moreover, by the Bellman equations (4.5) and by the equilibrium conditions (4.1), (4.2), (4.3) and (4.4) we have that x', ℓ' fulfill the following definition[2].

[2]This definition can be found also in Definition 4.6, in Chapter 4 of this thesis.

Definition 7.1 (Normalized Thin Flow with Resetting (NTF)[Cominetti et al., 2015])**.** *We say that the pair (x', ℓ') is an Normalized Thin Flow on $G_\theta = (V, E'_\theta, E^*_\theta)$ if:*

$$\begin{aligned}
&x' \text{ is a static flow on } (V, E'_\theta) \text{ with value } \nu_0 \\
&\ell'_s = 1 \\
&\ell'_w = \min_{vw \in E'_\theta} \eta(\ell'_v, x'_{vw}) && \forall v \in V \setminus \{s\} \\
&\ell'_w = \eta(\ell'_v, x'_{vw}) && \forall vw \in E'_\theta \text{ with } x'_{vw} > 0
\end{aligned}$$

where

$$\eta(\ell'_v, x'_{vw}) = \begin{cases} \frac{x'_{vw}}{\nu_{vw}} & vw \in E^*_\theta \\ \max\{\ell'_v, \frac{x'_{vw}}{\nu_{vw}}\} & vw \notin E^*_\theta. \end{cases}$$

Steady State. We say that a dynamic equilibrium is in a steady state at time θ if the shortest path network and the label derivatives $(\ell'_v)_{v \in V}$ do not change for any $\vartheta \in [\theta, \infty)$.

7.3 Characterization of steady states

In this section we give a characterization of the steady states. We first describe the derivatives of the equilibrium in a steady state and then we identify, and show how to compute the set of queue lengths that induce a steady state.

7.3.1 Derivatives of the equilibrium in a steady state

The derivatives of the equilibrium in a steady state are described through the following theorem:

Theorem 7.2. *The equilibrium is in a steady state at time θ if and only if the NTF on the shortest path network $G_\theta = (V, E'_\theta, E^*_\theta)$ equals a NTF on $(V, E, \varnothing)$.*

Proof. This proof consists of two parts, one for each direction.

For the first direction we need to show that, if the equilibrium is in a steady state at time θ, then the NTF (x', ℓ') on $G_\theta = (V, E'_\theta, E^*_\theta)$ corresponds to a NTF on $(V, E, \varnothing)$.

Since (x', ℓ') is a NTF on G_θ, we know that x' is a static flow on $G = (V, E)$ with value ν_0 and that $\ell'_s = 1$.

To show that

$$\ell'_w = \eta(\ell'_v, x'_{vw}) = \max\{\ell'_v, \frac{x'_{vw}}{\nu_{vw}}\} \qquad \forall vw \in E \text{ with } x'_{vw} > 0 \tag{7.1}$$

we need to show that $\ell'_w = \max\{\frac{x'_{vw}}{\nu_{vw}}, \ell'_v\}$ for all $vw \in E^*_\theta$. Consider an arc $e = vw \in E^*_\theta$ and a point in time $\vartheta \geq \theta$; since $\ell'_w = \frac{x'_{vw}}{\nu_{vw}}$ we just need to show that $\ell'_w \geq \ell'_v$. By definition we know that the resetting arcs are the arcs that host queues and that in a steady state

the rate of change of the transit time of a link does not change. As a consequence the queueing delay has a nonnegative rate of growth:

$$\frac{d}{d\vartheta}\Big(q_{vw}(\ell_v(\vartheta))\Big) \geq 0 \ .$$

By definition of shortest path network, $\ell_v(\vartheta) + \tau_e + q_e(\ell_v(\vartheta)) = \ell_w(\vartheta)$, and hence

$$\frac{d}{d\vartheta}\Big(q_{vw}(\ell_v(\vartheta))\Big) = \ell'_w - \ell'_v \ ,$$

implying that $\ell'_w \geq \ell'_v$.

Equation (7.1) and the fact that the flow on a resetting arc is positive imply that

$$\ell'_w = \min_{vw \in E'_\theta} \eta(\ell'_v, x'_{vw}) = \min_{vw \in E'_\theta} \max\{\ell'_v, \frac{x'_{vw}}{\nu_{vw}}\} \qquad \forall v \in V \setminus \{s\} \ .$$

Hence, to show that

$$\ell'_w = \min_{vw \in E} \eta(\ell'_v, x'_{vw}) = \min_{vw \in E} \max\{\ell'_v, \frac{x'_{vw}}{\nu_{vw}}\} \qquad \forall v \in V \setminus \{s\}$$

we just need to show that $\ell'_w \leq \ell'_v$ for all $vw \in E \setminus E'_\theta$. Consider an arc $e = vw \in E \setminus E'_\theta$; clearly $x'_e = 0$ and, by the steady state definition, $\ell_w(\vartheta) < \ell_v(\vartheta) + \tau_e$ for any $\vartheta \in [\theta, \infty)$, implying that $\ell'_w \leq \ell'_v$ and concluding the first part of the proof.

For the second direction we need to show that, if the NTF (x', ℓ') on $G_\theta = (V, E'_\theta, E^*_\theta)$ equals a NTF on $(V, E, \varnothing)$ then the equilibrium is in a steady state at time θ. Our goal is thus to show that the queues on E^*_θ never fully deplete and that an arc $e = vw \in E \setminus E'_\theta$ never joins the shortest path network.

The former holds by Definition 7.1 and since (x', ℓ') is a NTF on $(V, E, \varnothing)$, which imply that $\ell'_w - \ell'_v \geq 0$ for all vw where $x'_{vw} > 0$. This, in turn, implies that $\frac{d}{d\theta}\Big(q_{vw}(\ell_v(\theta))\Big) = \ell'_w - \ell'_v \geq 0$.

Consider now an arc $e = vw \in E \setminus E'_\theta$. Since $e \notin E'_\theta$ implies that $\ell'_w \leq \ell'_v$ (by Definition 7.1) and $\ell_w(\theta) < \ell_v(\theta) + \tau_e$ (by definition of shortest path network), the arc e never joins the shortest path network. □

7.3.2 Queue lengths of a steady state

In the following we characterize the possible queue lengths that can occur at some moment of a steady state.

Given a NTF (x', ℓ') on $(V, E, \varnothing)$, consider the graph $\bar{G} = (V, \bar{E})$ obtained from G by modifying each arc capacity as follows: the capacity of arc vw in $\bar{E}$ is $\bar{\nu}_{vw} = \nu_{vw} \cdot \ell'_w$. Let $(\mathbf{P})$ be a min-cost flow problem on $\bar{G}$ where the costs are represented by the transit time

$(\tau_e)_{e\in E}$ and let (**D**) be the dual of (**P**), i.e. let (**P**) and (**D**) be the following problems:

$$
\begin{aligned}
(\mathbf{P}):\quad \min\quad & \sum_{e\in E} \tau_e \cdot g_e \\
\text{s.t.}\quad & \sum_{e\in\delta^{out}(v)} g_e - \sum_{e\in\delta^{in}(v)} g_e = \begin{cases} 0 & \forall v \in V\setminus\{s,t\} \\ \nu_0 & \text{if } v = s \\ -\nu_0 & \text{if } v = t \end{cases} && (7.2)\\
& g_e \le \bar{\nu}_e \qquad \forall e \in E && (7.3)\\
& g_e \ge 0 \qquad \forall e \in E
\end{aligned}
$$

$$
\begin{aligned}
(\mathbf{D}):\quad \max\quad & \nu_0 d_t - \sum_{e\in E} \bar{\nu}_e q_e \\
s.t.\quad & d_w \le d_v + \tau_e + q_e \qquad \forall e = vw \in E && (7.4)\\
& q_e \ge 0 \qquad \forall e \in E
\end{aligned}
$$

We are going to show that an optimal solution to (**P**) corresponds to a NTF of a steady state and that an optimal solution to (**D**) provides us the queue lengths to assign to the arcs of the network to obtain a steady state.

Let g^* be an optimal solution to (**P**) and (d^*, q^*) be an optimal solution to (**D**). Let $G^+ = (V, E^+)$ where $E^+ = \{e \in E : g^*_e > 0\}$. Finally, recall that we defined (x', ℓ') to be an NTF on $(V, E, \varnothing)$. For the remainder of the section our goal is to prove the following theorem.

Theorem 7.3. *If, for a fixed θ, the waiting time $q_e(\ell_v(\theta)) = q^*_e$ for each $e \in E^+$, then the equilibrium is in a steady state at time θ and (g^*, ℓ') is a NTF of the steady state.*

To prove Theorem 7.3 we will need some supporting lemmas.

Lemma 7.4. *$\ell'_t = \max\{\ell'_v : v \in V\}$ and $\ell'_s = \min\{\ell'_v : v \in V\}$*

Proof. Consider an arbitrary s-t-path in G that uses only arcs in E^+. The statement holds since $\ell'_w \ge \ell'_v$ for any arc $vw \in E^+$. □

Lemma 7.5. *For any arc $e = vw \in E$, if $\ell'_v \ne \ell'_w$ then $x'_{vw} = g^*_{vw}$.*

Proof. Consider an arc $vw \in E$ such that $\ell'_v \ne \ell'_w$; let $z = \min\{\ell'_v, \ell'_w\}$ and let V_z be the set of vertices with labels ℓ' not greater than z, i.e. $V_z = \{u \in V : \ell'_u \le z\}$. Note that, by Lemma 7.4, we have that V_z contains s and does not contain t. Let $\delta^{out}(S)$ and $\delta^{in}(S)$ be the set of arcs in E with, respectively, tail and head in $S \subset V$. Since (x', ℓ') is a NTF on $(V, E, \varnothing)$, by Definition 7.1 we have that the total amount of flow x' entering V_z is zero and the total amount leaving it, denoted as $x'(\delta^{out}(V_z))$, satisfy the following equations:

$$x'(\delta^{out}(V_z)) = \sum_{e=vw\in E:\ell'_v\le z<\ell'_w} \nu_e \cdot \ell'_w \qquad (7.5)$$

and

$$x'(\delta^{out}(V_z)) = \nu_0\ . \qquad (7.6)$$

It follows that:

$$\begin{aligned} g^*(\delta^{out}(V_z)) &= \sum_{e=vw\in E:\ell'_v\le z<\ell'_w} g^*_e \\ &\le \sum_{e=vw\in E:\ell'_v\le z<\ell'_w} \bar{\nu}_e && \text{by (7.3)} \\ &= \sum_{e=vw\in E:\ell'_v\le z<\ell'_w} \nu_e\cdot\ell'_w && \text{by definition of } \bar{\nu}_e \\ &= \nu_0 && \text{by (7.5) and (7.6).} \end{aligned}$$

On the other hand, by Equation (7.2), $g^*(\delta^{in}(V\setminus V_z)) \ge \nu_0$ and, therefore

$$g^*(\delta^{out}(V_z)) = \nu_0\ .$$

This equality, together with Equation (7.2), implies that g^* sends no flow inside V_z and that $g^*_{vw} = \bar{\nu}_{vw} = \nu_{vw}\cdot\ell'_w = x'_{vw}$ for every $vw\in\delta^{out}(V_z)$, concluding the proof. □

Corollary 7.6. *(g^*,ℓ') is a NTF on $(V,E,\varnothing)$.*

Proof. In order to prove the statement we have to show that the conditions of Definition 7.1 hold. By Lemma 7.5, we just need to show that these are not violated by an arc $e=vw\in E$ with $\ell'_v=\ell'_w$. But this never happens since $g^*_e\le\bar{\nu}_e=\nu_e\cdot\ell'_w$, which implies that $\ell'_w=\max\{\frac{g^*_e}{\nu_e},\ell'_v\}$. □

With the next lemma we show that, if we assign to the arcs the queue lengths q^*, then g^* is positive only along the shortest paths of the network.

Lemma 7.7. *If, for a fixed θ, the waiting time $q_e(\ell_v(\theta)) = q^*_e$ for each $e\in E$, then $\ell_v(\theta)=d^*_v$ for any $v\in V$ and g^*_e is positive only if the arc e belongs to a shortest path.*

Proof. First of all notice that the transit time of any s-v-path in G^+ with queue waiting time q^* is equal to d^*_v: by complementary slackness, $g^*_e>0$ implies $d^*_w=d^*_v+\tau_e+q^*_e$ and, as a consequence, the transit time of any s-v-path in G^+ is equal to d^*_v.

Our goal is thus to show that if a s-v-path contains arcs in $E\setminus E^+$ then its transit time is not strictly smaller than d^*_v.

Let's proceed by contradiction and consider a shortest s-t-path in G which contains a v_1-v_2-subpath P_1 consisting only of arcs in $E\setminus E^+$ and with a smaller transit time than the one of a v_1-v_2-path P_2 in G^+. If we modify g^* by redirecting a positive amount of flow from P_2 into P_1, i.e. by reducing the flow on P_2 and increasing the flow on P_1 by $\varepsilon>0$, we obtain a solution to (**P**) that has a smaller cost than g^* (the transit time represents the cost in (**P**)), a contradiction to the optimality of g^*. □

We are now ready to prove Theorem 7.3.

Proof of Theorem 7.3. Let E^* be the set of arcs where q^* is positive, i.e. $E^*=\{e\in E: q^*_e>0\}$. Using Lemma 7.7 and Theorem 7.2, we just need to show that (g^*,ℓ') is a NTF on $G_\theta=(V,E,E^*)$ and, using Corollary 7.6 and Definition 7.1, we just need to show that $\ell'_w=\frac{g^*_{vw}}{\nu_{vw}}$ for all $vw\in E^*$. But this holds by complementary slackness: $q^*_e>0$ implies $g^*_e=\bar{\nu}_e=\nu_e\cdot\ell'_w$ for any arc $e=vw\in E^*$. □

7.4 Candidate potential function

Let $(\hat{x}', \hat{\ell}')$ be a NTF on $(V, E, \varnothing)$ and consider the following potential function

$$\Phi(\theta) = \nu_0(\ell_t(\theta) - \ell_s(\theta)) - \sum_{e=vw \in E} \nu_e \cdot \hat{\ell}'_w \cdot q_e(\ell_v(\theta)) \ .$$

Note that $\Phi(\theta)$ is inspired from the objective of the previous dual program (**D**) and generalizes the potential function used in [Cominetti et al., 2017], where $\hat{\ell}'_w = 1$ for any $w \in V$.

With the following lemmas we show that $\Phi(\theta)$ is bounded and that $\Phi'(\theta) = 0$ if the equilibrium is in a steady state at time θ.

Lemma 7.8. *$\Phi(\theta)$ is bounded.*

Proof. First of all, (**P**) is clearly bounded and it always admits a solution since an optimal one corresponds to a NTF on $(V, E, \varnothing)$ (Corollary 7.6) and this always exists [Cominetti et al., 2015]. Then, from the earliest arrival functions definition (Equation (4.5)), the point (d, q) with $d_v = \ell_v(\theta)$ for all $v \in V$ and $q_e = q_e(\ell_v(\theta))$ is feasible for the dual (**D**). Therefore $\Phi(\theta)$ corresponds to the value of a feasible solution of (**D**) and is bounded by the optimal value of (**P**). □

Lemma 7.9. *For a NTF on $(V, E, \varnothing)$, $\Phi'(\theta) = 0$.*

Proof. Let $E^+ = \{e \in E : x'_e > 0\}$ and let (x', ℓ') be a NTF on $(V, E, \varnothing)$. Let $\mathcal{P}$ be the set of all simple s-t-paths and consider a path-decomposition $(x'_P)_{P \in \mathcal{P}}$ of x' [3]. Then:

$$\begin{aligned}
\Phi'(\theta) &= \nu_0(\ell'_t - \ell'_s) - \sum_{e=vw \in E^+} \nu_e \hat{\ell}'_w (\ell'_w - \ell'_v) \\
&= \nu_0(\ell'_t - \ell'_s) - \sum_{e=vw \in E^+} \nu_e \ell'_w (\ell'_w - \ell'_v) \\
&= \nu_0(\ell'_t - \ell'_s) - \sum_{e=vw \in E^+ : \ell_w \neq \ell_v} x'_e (\ell'_w - \ell'_v) && \text{by Definition 7.1} \\
&= \nu_0(\ell'_t - \ell'_s) - \sum_{P \in \mathcal{P}} \sum_{e=vw \in P : \ell_w \neq \ell_v} x'_P (\ell'_w - \ell'_v) \\
&= \nu_0(\ell'_t - \ell'_s) - \sum_{P \in \mathcal{P}} x'_P (\ell'_t - \ell'_s) \\
&= \nu_0(\ell'_t - \ell'_s) - \nu_0(\ell'_t - \ell'_s) \\
&= 0 \ .
\end{aligned}$$

□

Given that $\Phi'(\theta)$ can take on only a finite number of values, depending on the NTF and more precisely on the current shortest path network, if $\Phi(\theta)$ were strictly increasing

[3] A path-decomposition $(x_P)_{P \in \mathcal{P}}$ of x is a collection of flow such that

$$x_P \geq 0 \ \forall P \in \mathcal{P} \quad \text{and} \quad x_e = \sum_{P \in \mathcal{P} : e \in P} x_P$$

where $e \in P$ indicates if path P contains that arc $e \in E$.

for any θ where the equilibrium is not in a steady state, then we would know that the equilibrium reaches a steady state in finite time. Unfortunately, as shown in the next example, we found a network and a queue lengths configuration in which the potential function decreases. However, we don't have an example of an evolution starting from an initially empty network where Φ' becomes negative and we suspect that it never does.

Example 7.1. Consider the graph $G = (V, E)$ of Figure 7.1, where

$$V = \{s, t, a_1, a_2, \ldots, a_{10}, b_1, b_2, \ldots, b_{15}\}$$

and

$$E = \{a_i a_{i+1} : i \in [9]\} \cup \{b_i b_{i+1} : i \in [14]\} \cup \{sa_1, a_{10}b_1, b_{15}t\} \cup \{sb_1, sb_{15}, a_1 t\} \ ,$$

with capacities:

$$\nu_e = \begin{cases} 2^{-4} & \text{if } e = sa_1 \\ 2^{-4-i} & \text{if } e = a_i a_{i+1}, i \in [9] \\ 2^{-14} & \text{if } e = a_{10}b_1 \\ 2^{-14+i} & \text{if } e = b_i b_{i+1}, i \in [14] \\ 2^{-2} & \text{if } e = b_{15}t \\ \infty & \text{if } e = sb_1 \\ \infty & \text{if } e = sb_{15} \\ \infty & \text{if } e = a_1 t \ . \end{cases}$$

We remark that there are smaller instances where the change in potential in negative but we chose this graph for numerical convenience.

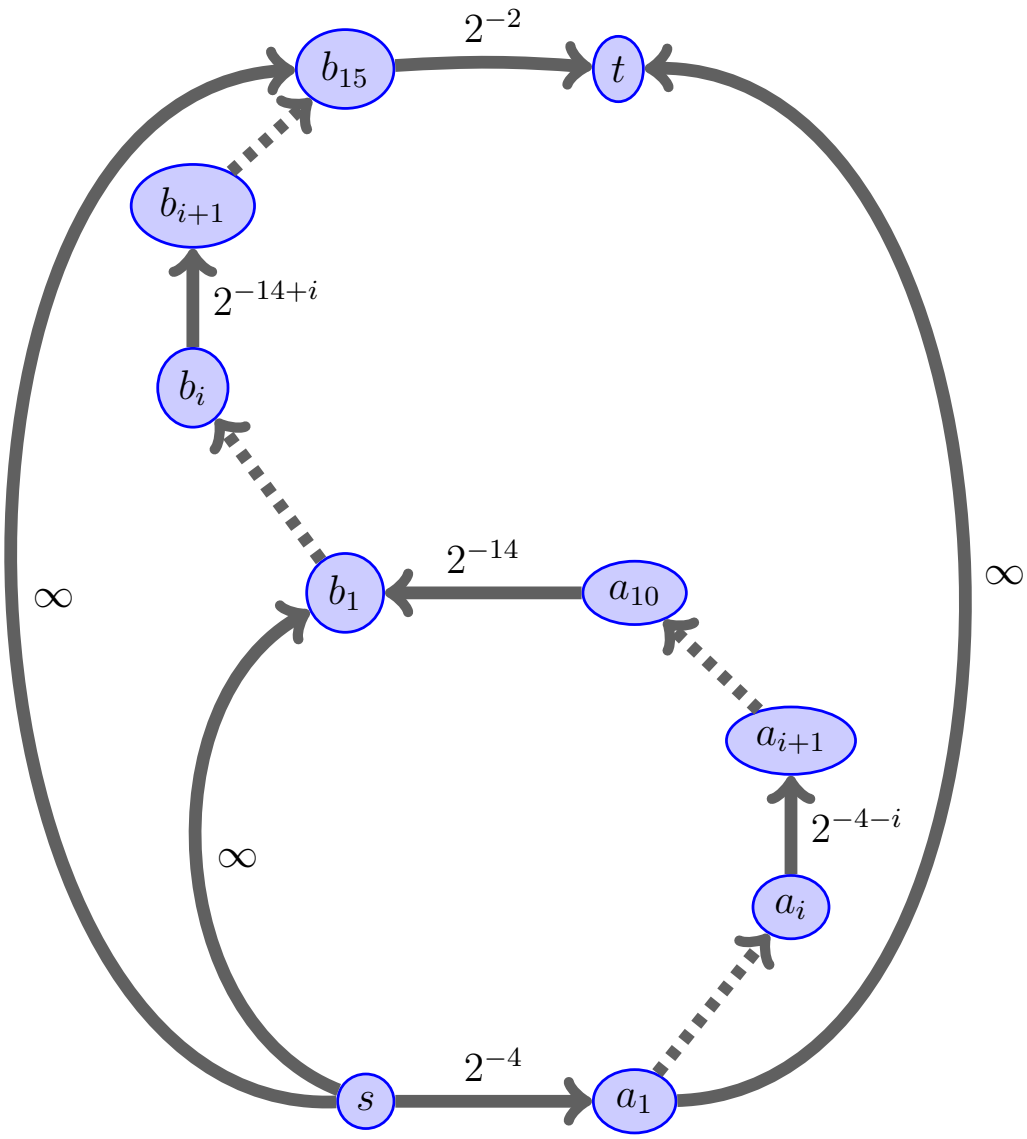


Figure 7.1: Graph of Example 7.1. The dash arcs represent paths and the label of an arc indicates the capacity.

The minimum s-t cut is $\{sa_1, b_{15}t\}$ and thus, for a NTF $(\hat{x}', \hat{\ell}')$ on $(V, E, \varnothing)$ with inflow $\nu_0 = 1$, the label $\hat{\ell}'_v$ is

$$\hat{\ell}'_v = \begin{cases} 3.2 & \text{if } v \in \{a_1, \dots, a_{10}, t\} \\ 1 & \text{if } v \in \{s, b_1, \dots, b_{15}\} \end{cases}.$$

Let $\bar{E}$ be the subset of E containing all the arcs with finite capacity, namely $\bar{E} = \{a_i a_{i+1} : i \in [9]\} \cup \{b_i b_{i+1} : i \in [14]\} \cup \{sa_1, a_{10}b_1, b_{15}t\}$. If at some point θ the current shortest path network contains all and only the arcs in $\bar{E}$ and these are all resetting, formally if $G_\theta = (V, \bar{E}, \bar{E})$, then the change in potential is negative: an NTF (x', ℓ') on $(V, \bar{E}, \bar{E})$ sends flow only along the arcs of $\bar{E}$ (see Figure 7.2), and consequently we have that

$$\begin{aligned}
\Phi' &= \nu_0(\ell'_t - \ell'_s) - \sum_{e=vw \in \bar{E}} \nu_e \hat{\ell}'_w(\ell'_w - \ell'_v) \\
&= 1(2^2 - 1) - 3.2 \cdot 2^{-4}(2^4 - 1) - 2^{-14}(2^{14} - 2^{13}) - 3.2 \sum_{i=5}^{13} 2^{-i}(2^i - 2^{i-1}) \\
&\quad - \sum_{i=2}^{15} 2^{-15+i}(2^{15-i} - 2^{16-i}) \\
&= 0 - \frac{1}{2} - 3.2 \sum_{i=5}^{13}(1 - \frac{2^{i-1}}{2^i}) - \sum_{i=2}^{15}(1 - \frac{2^{16-i}}{2^{15-i}}) \\
&= -3.2 \cdot 9 \cdot \frac{1}{2} - \frac{1}{2} + 14 \\
&= -0.9.
\end{aligned}$$

Notice that, in this specific instance, the evolution starting from an empty network never reaches a phase where Φ' is negative.

7.5 Conclusion

This is the only chapter in the thesis where users do not choose their departure time but are released at a constant inflow rate. As seen in Chapter 4, this setting is a special case of the more general setting that has been treated in the rest of the thesis; for this reason, understanding the structure of the equilibria in this framework would provide important insights to the endogenous one.

Here, we studied the steady states of equilibria when the inflow rate into the network exceed the capacity of the minimum s-t-cut. We provided a full characterization of the steady states and we explore a technique to verify whether an equilibrium always reach such a state (as in the case where the inflow rate is bounded by the capacity of the minimum s-t-cut).

Unfortunately our technique did not bring the hoped results: we found a network and queue lengths configuration where it fails. However, we suspect that steady states are always attained and that a deeper understanding of the relationship between consecutive thin flows with resetting is the key to establish it.

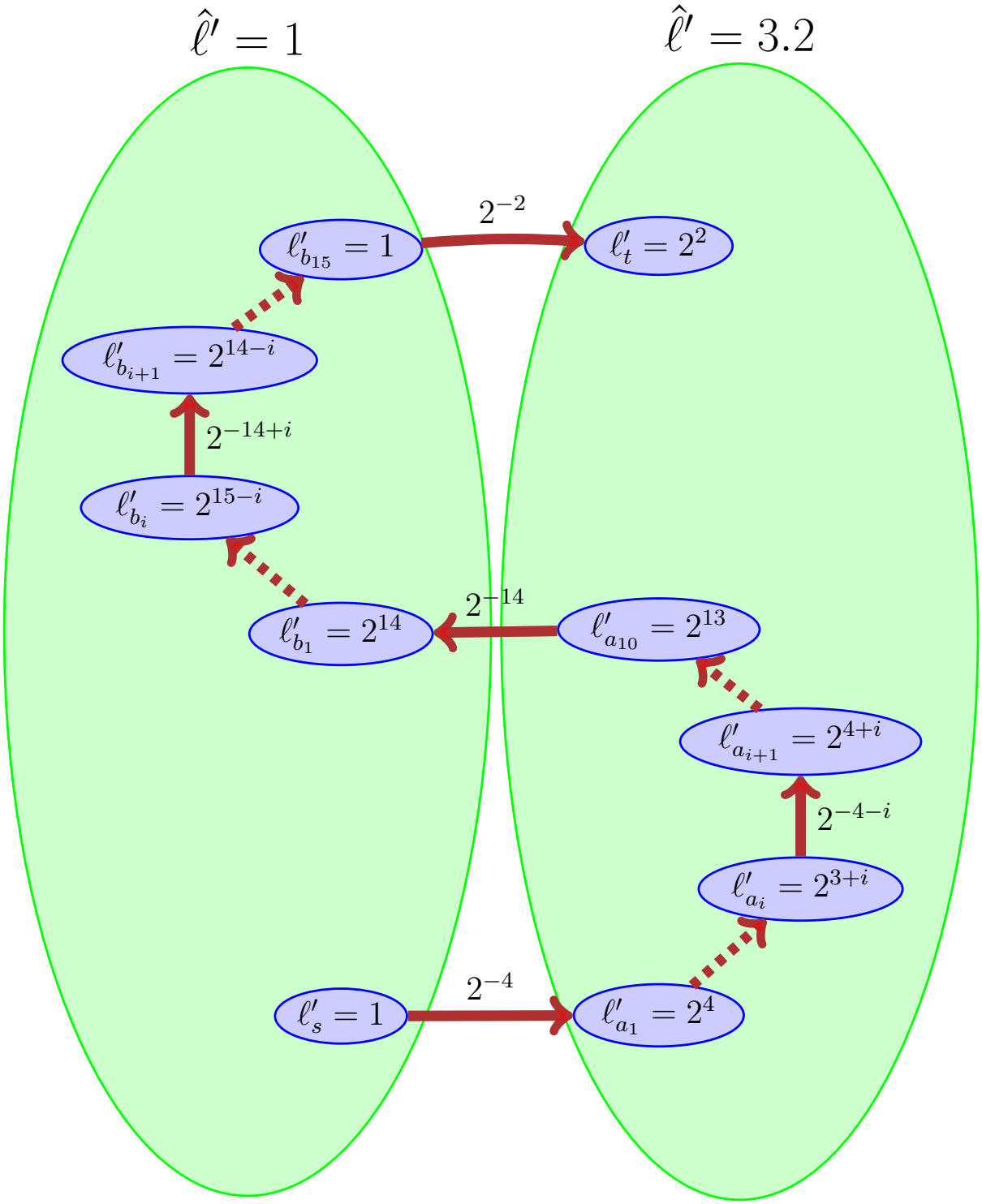


Figure 7.2: Example of a NTF for which $\Phi' < 0$. The label of an arc represents the capacity. The dash arcs represent paths.

Summary and Conclusion

In this thesis we made significant progress in understanding the structure of dynamic traffic equilibria.

One of the main contributions has been introducing (in Chapter 4), using methodologies from the optimization literature, a framework to study equilibria on general networks where agents have route and departure time choices. This is of interest not only for the optimization community, that considered similar models but only with route choices, but also for the transportation economics community, that considered similar models but on simple networks. Moreover, this model is not only interesting from an academic point of view but can also be implemented by network designers to run simulations for real-world scenarios. Indeed, this framework provides compelling insights into the dynamics of the equilibrium behavior; it allows us to keep track of how much traffic is on each link of the network per any given time even when the number of vehicles is massive.

A clear example of this is given in Chapter 6, when it is used to provide a clear explicit characterization of the evolution behavior in particular network topologies, the ones of the corridor model.

Furthermore, we showed that this framework can represent empirically observed behaviors like the one of hypercongestion (Chapter 5), one of the most severe types of traffic congestion. Identifying the causes of hypercongestion is extremely valuable since it can help in the design of better road networks or in the formulation of better policies that partially or fully prevent congestion.

Regarding policies, we showed, in Chapter 3, how to compute tolls that induce an optimal equilibrium – the equilibrium that maximizes the social welfare – for both the setting with and without departure time choice.

Some fundamental questions in both the settings with and without departure time choice still need to be settled. For example, some technical issues in the proofs of existence and uniqueness of equilibria still need to be resolved. Another unsolved fundamental query is the one regarding the complexity of the thin flow with resetting and the complexity of computing the entire equilibrium behavior. Regarding the latter we do not even know yet if it can be computed in finite time. We believe that the crux of these latter questions, as well as of the long term behavior studied in Chapter 7, lies on the dependencies between the thin flows and the sets of active and resetting arcs of consecutive shortest path networks.

In addition to these open problems there are many possible research directions. For example, in this thesis we considered homogeneous agents and a single origin-destination pair. It would be very interesting to extend our study to settings with user heterogeneity, for example in the value of time, in scheduling preferences or in the origin-destination pairs.

Other possible directions would be to investigate Braess's paradox on our framework or to consider a framework similar to ours but with uncertainty in the delays experienced on links.

Acknowledgments

This thesis is the result of the work I did over four years at the Vrije Universiteit Amsterdam and I would like to thank all the people that helped me along the way.

First and foremost, I would like to thank my supervisor, Neil Olver, for the time he invested in me, for our fruitful discussions and for his constructive feedback. He was available all the time to answer my questions and always provided me with insightful comments. His knowledge and passion have contributed enormously to all the results contained in this thesis. Working with him has been very instructive, I learned and grew a lot and I will always be grateful for this.

I would like to thank my other supervisor, Leen Stougie. I truly enjoyed all our talks and always appreciated his enthusiasm about research and collaboration.

A big thank goes to Erik Verhoef without whose contributions the thesis would have not be in its current shape. After our first collaboration, I got inspired and invested more time on dynamic equilibria with departure time choice and eventually this became the main subject of the thesis.

Many thanks go to the rest of the thesis committee: Tobias Harks, Tim Oosterwijk, Britta Peis, Jan Rouwendal, Guido Schäfer and Vincent Van den Berg. I thank them all for taking the time to review my thesis and for their valuable comments.

A special thanks goes to Gabriel Moruz, who was my Master's thesis advisor during my time in Frankfurt Am Main. He guided me throughout my first theoretical research project and he motivated me to pursue a PhD. If it wasn't for him, I might have taken a different career path.

I also want to thank all the people I have worked with during this period: Thomas Bosman, Andrés Cristi, Marcus Kaiser, Tim Oosterwijk, Leon Sering, René Sitters and Laura Vargas Koch. It was a pleasure working with them.

Thanks to the Networks and Optimization Group at Centrum Wiskunde & Informatica for hosting me for years. It has been a really pleasant working environment and I am glad I had the opportunities to be there.

Thanks to all the anonymous reviewers who helped improving this work and thanks to the academic community as a whole which has provided me a lot of free knowledge.

Finally, I want to thank my parents, without whom I would have not make it this far.

Bibliography

R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993.

T. Akamatsu, K. Wada, and S. Hayashi. The corridor problem with discrete multiple bottlenecks. *Transportation Research Part B: Methodological*, 81:808–829, 2015.

E. Anshelevich and S. Ukkusuri. Equilibria in dynamic selfish routing. In *International Symposium on Algorithmic Game Theory*, pages 171–182, 2009.

R. Arnott. The corridor problem. Working Papers in Economics, 2001.

R. Arnott. A bathtub model of downtown traffic congestion. *Journal of Urban Economics*, 76:110–121, 2013.

R. Arnott and E. DePalma. The corridor problem: preliminary results on the no-toll equilibrium. *Transportation Research Part B: Methodological*, 45(5):743–768, 2011a.

R. Arnott and E. DePalma. The corridor problem: Preliminary results on the no-toll equilibrium. *Transportation Research Part B: Methodological*, 45(5):743–768, 2011b. doi: 10.1016/j.trb.2011.01.004.

R. Arnott, A. de Palma, and R. Lindsey. Economics of a bottleneck. *Journal of Urban Economics*, 27(1):111–130, 1990.

R. Arnott, A. de Palma, and R. Lindsey. Properties of dynamic traffic equilibrium involving bottlenecks, including a paradox and metering. *Transportation Science*, 27:148–160, 1993.

R. Arnott, A. Kokoza, and M. Naji. Equilibrium traffic dynamics in a bathtub model: A special case. *Economics of Transportation*, 7-8:38–52, 2016.

N. Baumann and M. Skutella. Solving evacuation problems efficiently–earliest arrival flows with multiple sources. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 399–410, 2006.

U. Bhaskar, L. Fleischer, and E. Anshelevich. A Stackelberg strategy for routing flow over time. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, (SODA)*, pages 192–201, 2011.

D. Braess. Über ein paradoxon aus der verkehrsplanung. *Unternehmensforschung*, 12(1): 258–268, 1968.

Z. Cao, B. Chen, X. Chen, and C. Wang. A network game of dynamic traffic. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 695–696, 2017.

X. Chu. Endogenous trip scheduling: The Henderson approach reformulated and compared with the Vickrey approach. *Journal of Urban Economics*, 37(3):324–343, 1995.

R. Cominetti, J. Correa, and O. Larré. Dynamic equilibria in fluid queueing networks. *Operations Research*, 63(1):21–34, 2015.

R. Cominetti, J. R. Correa, and N. Olver. Long term behavior of dynamic equilibria in fluid queuing networks. In *Integer Programming and Combinatorial Optimization (IPCO)*, pages 161–172, 2017.

J. R. Correa, A. Cristi, and T. Oosterwijk. On the price of anarchy for flows over time. In *In Proceedings of the 2019 ACM Conference on Economics and Computation, (EC)*, pages 559–577, 2019.

C. F. Daganzo, V. V. Gayah, and E. J. Gonzales. Macroscopic relations of urban traffic variables: Bifurcations, multivaluedness and instability. *Transportation Research Part B: Methodological*, 45(1):278–288, 2011.

A. De Palma, M. Kilani, and R. Lindsey. Comparison of second-best and third-best tolling schemes on a road network. *Transportation Research Record*, 1932(1):89–96, 2005.

Y. Disser and M. Skutella. The simplex algorithm is NP-mighty. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, (SODA)*, pages 858–872, 2015.

L. Fleischer and E. Tardos. Efficient continuous-time dynamic network flow algorithms. *Operations Research Letters*, 23(3):71–80, 1998.

L. R. Ford and D. R. Fulkerson. Constructing maximal dynamic flows from static flows. *Operations Research*, 6(3):419–433, 1958.

L. R. Ford and D. R. Fulkerson. *Flows in Networks*. Princeton University Press, 1962.

M. Fosgerau. Congestion in the bathtub. *Economics of Transportation*, 4(4):241–255, 2015.

D. Frascaria and N. Olver. Algorithms for flows over time with scheduling costs. In *Integer Programming and Combinatorial Optimization (IPCO)*, pages 130–143, 2020.

D. Frascaria, N. Olver, and E. Verhoef. Emergent hypercongestion in Vickrey bottleneck networks. *Transportation Research Part B: Methodological*, 139:523 – 538, 2020.

T. L. Friesz and K. Han. The mathematical foundations of dynamic user equilibrium. *Transportation Research Part B: Methodological*, 126:309–328, 2019.

T. L. Friesz, D. Bernstein, T. E. Smith, R. L. Tobin, and B.-W. Wie. A variational inequality formulation of the dynamic network user equilibrium problem. *Operations research*, 41(1):179–191, 1993.

D. Gale. Transient flows in networks. *The Michigan Mathematical Journal*, 6(1):59–63, 1959.

N. Geroliminis and C. F. Daganzo. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, 42(9):759–770, 2008.

L. Graf and T. Harks. Dynamic flows with adaptive route choice. In *Integer Programming and Combinatorial Optimization (IPCO)*, volume 11480 of *Lecture Notes in Computer Science*, pages 219–232. Springer, 2019.

L. Graf and T. Harks. A finite time combinatorial algorithm for instantaneous dynamic equilibrium flows. Preprint, arXiv:2007.07808, 2020a. URL https://arxiv.org/abs/2007.07808.

L. Graf and T. Harks. The price of anarchy for instantaneous dynamic equilibria. Preprint, arXiv:2007.07794, 2020b. URL https://arxiv.org/abs/2007.07794.

L. Graf, T. Harks, and L. Sering. Dynamic flows with adaptive route choice. *Mathematical Programming*, 183(1):309–335, 2020.

B. D. Greenshields. A study of traffic capacity. In *Highway research board proceedings*, volume 14, pages 448–477, 1935.

T. Harks. Pricing in resource allocation games based on duality gaps. Preprint, arXiv:1907.01976, 2019. URL http://arxiv.org/abs/1907.01976.

M. Hoefer, V. S. Mirrokni, H. Röglin, and S.-H. Teng. Competitive routing over time. *Theoretical Computer Science*, 412(39):5420–5432, 2011.

A. Ismaili. Routing games over time with fifo policy. In *International Conference on Web and Internet Economics*, pages 266–280, 2017.

B. N. Janson. Dynamic traffic assignment for urban road networks. *Transportation Research Part B: Methodological*, 25(2-3):143–161, 1991.

J. J. Jarvis and H. D. Ratliff. Some equivalent objectives for dynamic network flow problems. *Management Science*, 28(1):106–109, 1982.

M. Kaiser. Computation of dynamic equilibria in series-parallel networks. Preprint, arXiv:2002.11428, 2020. URL http://arxiv.org/abs/2002.11428.

F. H. Knight. Some fallacies in the interpretation of social cost. *The Quarterly Journal of Economics*, 38(4):582–606, 1924.

R. Koch and M. Skutella. Nash equilibria and the price of anarchy for flows over time. *Theory of Computing Systems*, 49(1):71–97, 2011.

E. Köhler, R. H. Möhring, and M. Skutella. Traffic networks and flows over time. In *Algorithmics of Large and Complex Networks - Design, Analysis, and Simulation*, pages 166–196, 2009.

G. Kolata. What if they closed 42d street and nobody noticed? *New York Times*, 1990.

E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. In *Annual Symposium on Theoretical Aspects of Computer Science*, pages 404–413, 1999.

M. Kuwahara. Equilibrium queueing patterns at a two-tandem bottleneck during the morning peak. *Transportation Science*, 24:217–229, 1990.

Z.-C. Li, H.-J. Huang, and H. Yang. Fifty years of the bottleneck model: A bibliometric review and future research directions. *Transportation research part B: methodological*, 139:311–342, 2020.

M. J. Lighthill and G. B. Whitham. On kinematic waves ii. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 229(1178):317–345, 1955.

C. R. Lindsey and E. T. Verhoef. Congestion modelling. Technical report, Tinbergen Institute Discussion Paper, 1999.

R. Lindsney and E. Verhoef. *Traffic congestion and congestion pricing*. Emerald Group Publishing Limited, 2001.

M. Macko, K. Larson, and L. Steskal. Braess's paradox for flows over time. In *International Symposium on Algorithmic Game Theory*, pages 262–275, 2010.

Q. Meng, W. Xu, and H. Yang. Trial-and-error procedure for implementing a road-pricing scheme. *Transportation research record*, 1923(1):103–109, 2005.

E. Minieka. Maximal, lexicographic, and dynamic network flows. *Operations Research*, 21 (2):517–527, 1973.

J. Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951.

N. Nisan, T. Roughgarden, É. Tardos, and V. V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge University Press, 2007. ISBN 9780511800481.

C. Papadimitriou. Algorithms, games, and the internet. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 749–753, 2001.

S. Peeta and A. K. Ziliaskopoulos. Foundations of dynamic traffic assignment: The past, the present and the future. *Networks and Spatial Economics*, 1(3):233–265, 2001.

A. B. Philpott. Continuous-time flows in networks. *Mathematics of Operations Research*, 15(4):640–661, 1990.

A. Pigou. *The Economics of Welfare*. Macmillan, 1920.

T. W. Reiland. Optimality conditions and duality in continuous programming ii. the linear problem revisited. *Journal of Mathematical Analysis and Applications*, 77(2):329–343, 1980. ISSN 0022247X. doi: 10.1016/0022-247X(80)90230-9.

P. I. Richards. Shock waves on the highway. *Operations research*, 4(1):42–51, 1956.

H. E. Romeijn, R. L. Smith, and J. C. Bean. Duality in infinite dimensional linear programming. *Mathematical Programming*, 53:79–97, 1992.

M. Scarsini, M. Schröder, and T. Tomala. Dynamic atomic congestion games with seasonal flows. *Operations Research*, 66(2):327–339, 2018.

L. Sering and M. Skutella. Multi-source multi-sink Nash flows over time. In *18th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems, (ATMOS)*, pages 12:1–12:20, 2018.

L. Sering and L. Vargas Koch. Nash flows over time with spillback. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, (SODA)*, pages 935–945, 2019.

T. C. Sharkey. *Infinite Linear Programs*. Wiley, 2011. doi: 10.1002/9780470400531.eorms0404. URL http://onlinelibrary.wiley.com/doi/abs/10.1002/9780470400531.eorms0404.

M. Skutella. An introduction to network flows over time. In *Research Trends in Combinatorial Optimization*, pages 451–482, 2009.

K. A. Small. The bottleneck model: An assessment and interpretation. *Economics of Transportation*, 4(1):110–117, 2015.

K. A. Small and E. Verhoef. *The Economics of Urban Transportation*. Routledge, 2007.

V. van den Berg and E. T. Verhoef. Winning or losing from dynamic bottleneck congestion pricing? the distributional effects of road pricing with heterogeneity in values of time and schedule delay. *Journal of Public Economics*, 95(7):983–992, 2011.

E. T. Verhoef. An integrated dynamic model of road traffic congestion based on simple car-following theory: Exploring hypercongestion. *Journal of Urban Economics*, 49(3): 505–542, 2001.

E. T. Verhoef. Inside the queue: Hypercongestion and road pricing in a continuous time–continuous place model of traffic congestion. *Journal of Urban Economics*, 54 (3):531–565, 2003.

W. Vickrey. Congestion theory and transport investment. *American Economic Review*, 59(2):251–60, 1969.

W. S. Vickrey. Pricing in urban and suburban transport. *The American Economic Review*, 53(2):452–465, 1963.

A. A. Walters. The theory and measurement of private and social cost of highway congestion. *Econometrica: Journal of the Econometric Society*, pages 676–699, 1961.

J. G. Wardrop. Road paper. some theoretical aspects of road traffic research. *Proceedings of the institution of civil engineers*, 1(3):325–362, 1952.

T. Werth, M. Holzhauser, and S. O. Krumke. Atomic routing in a deterministic queuing model. *Operations Research Perspectives*, 1(1):18–41, 2014.

W. L. Wilkinson. An algorithm for universal maximal dynamic flows in a network. *Operations Research*, 19(7):1602–1612, 1971.

H. Yang and Q. Meng. Departure time, route choice and congestion toll in a queuing network with elastic demand. *Transportation Research Part B: Methodological*, 32(4): 247–260, 1998. ISSN 0191-2615.

H. Yang, Q. Meng, and D.-H. Lee. Trial-and-error implementation of marginal-cost pricing on networks in the absence of demand functions. *Transportation Research Part B: Methodological*, 38(6):477–493, 2004.

N. Zadeh. A bad network problem for the simplex method and other minimum cost flow algorithms. *Mathematical Programming*, 5:255–266, 1973.

X. Zhang, W. Lam, and H.-J. Huang. Braess's paradoxes in dynamic traffic assignment with simultaneous departure time and route choices. *Transportmetrica*, 4:209–225, 2008.